REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	1996	

4. TITLE AND SUBTITLE
Verification of Vortex '94 Forecasts

5. FUNDING NUMBERS

6. AUTHOR(S)

Robert d. Duncomb Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

AFIT Student Attending:
University of Oklahoma

8. PERFORMING ORGANIZATION REPORT NUMBER
96-036

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

DEPARTMENT OF THE AIR FORCE
AFIT/CI
2950 P STEET, BLDG 125
WRIGHT-PATTERSON AFB OH 45433-7765

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release IAW 190-1
Distribution Unlimited
BRIAN D. GAUTHIER, MSgt, USAF
Chief Administration

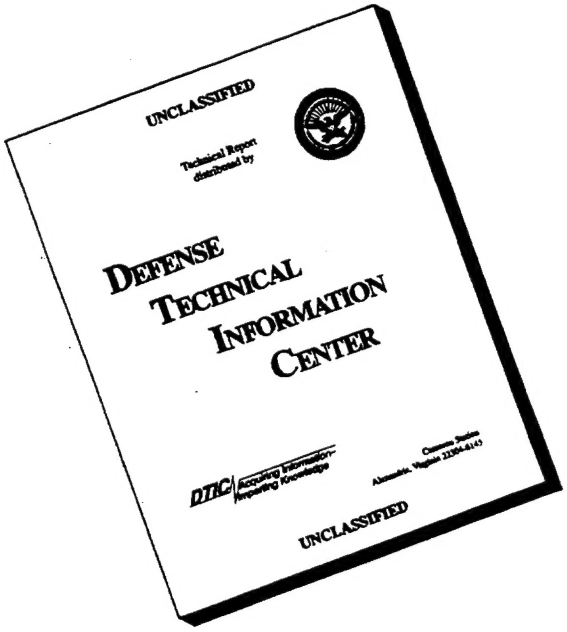
12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

14. SUBJECT TERMS	15. NUMBER OF PAGES 130
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to ***stay within the lines*** to meet ***optical scanning requirements***.

Block 1. Agency Use Only *(Leave blank)*.

Block 2. Report Date. Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least the year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C	- Contract	**PR**	- Project
G	- Grant	**TA**	- Task
PE	- Program Element	**WU**	- Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es). Self-explanatory.

Block 10. Sponsoring/Monitoring Agency Report Number. *(If known)*

Block 11. Supplementary Notes. Enter information not included elsewhere such as: Prepared in cooperation with...; Trans. of...; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

- **DOD** - See DoDD 5230.24, "Distribution Statements on Technical Documents."
- **DOE** - See authorities.
- **NASA** - See Handbook NHB 2200.2.
- **NTIS** - Leave blank.

Block 12b. Distribution Code.

- **DOD** - Leave blank.
- **DOE** - Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.
- **NASA** - Leave blank.
- **NTIS** - Leave blank.

Block 13. Abstract. Include a brief *(Maximum 200 words)* factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code *(NTIS only)*.

Blocks 17. - 19. Security Classifications. Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.

★U.S.GPO:1990-0-273-271

UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

VERIFICATION OF VORTEX '94 FORECASTS

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

MASTER OF SCIENCE IN METEOROLOGY

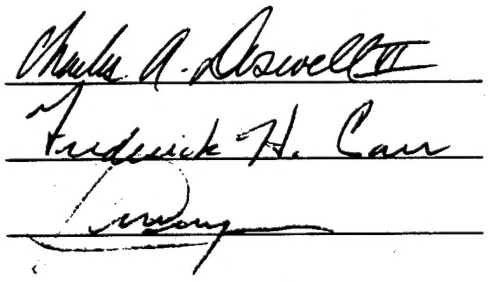
By

ROBERT D. DUNCOMB JR.
Norman, Oklahoma
1996

VERIFICATION OF VORTEX '94 FORECASTS

A THESIS APPROVED FOR THE
SCHOOL OF METEOROLOGY

BY



ACKNOWLEDGEMENTS

The author wishes to thank his advisor, Charles A. Doswell III, for his advice and patience through this project and for keeping the project on schedule in the short amount of time available to complete it.

The author also wishes to thank Dr. Fred Carr and Dr. Kelvin Droegemeir for their willingness to be on the advising committee. Their input challenged him to think about this project in new ways. Thanks are also due to the many people who helped the author get the data needed to complete this project including, at NWS/EFF OUN: Mike Branick; at NSSL: Erik Rasmussen, Phil Bothwell, Doug Rhue, and Irv Watson; and at NSSFC/SELS: John Halmstad and Jan Lewis.

Finally, the author is especially grateful to his family who were very understanding when he had to spend long hours on this project. To his wife, thank you for your support, encouragement, and love throughout. To his kids, Macy, Jonah, Jonathan, and Jeremiah, thanks for the time you had to sacrifice away from your father and for making him laugh after many of the stressful days.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

TABLE Page

TABLE Page

LIST OF ILLUSTRATIONS

FIGURE Page

FIGURE Page

FIGURE Page

Abstract

Forecast verification involves the assessing the quality of a set of forecasts. There are many different methods of measuring the quality of a set of forecasts, and this study examines two; the measures-oriented approach and the distributions-oriented approach. It is argued that the distributions-oriented approach offers a more comprehensive look at the set of forecasts and observations and the relationships between them. This approach also highlights the strengths and weaknesses of the forecasting system so that improvements can be made in specific areas.

This study evaluates the forecasts made during VORTEX '94. The forecasts included experimental probabilistic forecasts for the VORTEX forecast area as well as probability contour forecasts. These forecasts were evaluated by looking at the conditional distribution of the forecast given the event, $p(f \mid x)$, the conditional distribution of the event given the forecast, $p(x \mid f)$, as well as the marginal distributions of forecasts, $p(f)$, and observations, $p(x)$. It is only through looking at these distributions of forecasts and observations that the true quality of a set of forecasts can be evaluated.

CHAPTER 1

INTRODUCTION

1.1 Motivation

It is only through a thorough verification process that the strengths and weaknesses of a forecast system can be identified, and the forecasts can be improved in time. Conversely, without performing a verification, it's not known whether there has been an increase in forecast quality and/or improvements in the forecasting system itself. As Murphy (1991) points out, "failure to consider the verification of forecasts, may lead us to erroneous conclusions about how well the forecasting system is doing."

Feedback to the forecasters on specific areas they can improve is vital to improving the forecasts themselves. Unfortunately, forecasters often make forecasts for which there is no immediate feedback on the quality of those forecasts. Since the forecasts will never be perfect, there should be concern with knowing forecast quality; progress in weather prediction comes from the failures of the past. These failures lead researchers on a continuing quest to understand the errors. If these failures are ignored, computers may eventually replace human forecasters. There is hope, however, if forecasters are willing to learn from their mistakes. Forecast verification is essential if there is concern at all with the quality of the forecasts.

1.2 Developing a verification methodology

There is, as yet, no clear definition for defining a "good" forecast. In this paper, we choose to follow Murphy (1993), who defines three types of "goodness" in weather forecasting: **type 1** goodness is satisfied when a forecaster's judgment shows consistency with his issued forecast; **type 2** goodness is the quality of the forecast which looks at how closely the forecasts match the observations; **type 3** goodness is the value of the forecast to potential users. Forecast consistency is completely under the control of the forecaster but it is hard to evaluate, as it requires knowledge of the subjective decisions made by an individual in preparing his/her forecast. The value of the forecast to potential users is user-dependent and can't be evaluated objectively for all users. Hopefully, those performing the verification consider the potential users of the forecast when a methodology is developed that will measure the quality of the forecast. The only type of goodness that can be evaluated objectively in a forecast verification is the forecast quality, given as the correspondence between the forecasts and the observations (Murphy 1993).

The first operational forecasts were issued in many locations in the United States and Europe between 1850-1870. It was soon after this that the need for a verification methodology was recognized. One of the first verification studies was done by Finley in 1884 for a set of tornado forecasts. Murphy (1996) discusses Finley's study, as well as the papers that followed in response, in what Murphy terms

"The Finley Affair". Murphy's review of several papers from 1884-1893 reveals that many of our modern-day verification measures were developed during this period.

Most verification studies since Finley's time have taken a **measures-oriented approach** to forecast verification. This approach uses individual measures of accuracy (correspondence between forecasts and observations) and skill (improvement over and above some reference forecast, e.g, climatology) to describe forecast quality. For probability forecasts, the Brier score (Brier 1950) is typically used as a measure of forecast accuracy, while the skill score (Sanders 1963) is used as a measure of forecast skill. These two scores have been widely used in past studies to verify both operational and experimental **probability** forecasts including: NWS PoP forecasts (Dagostaro et al. 1995, Carter and Polger 1986); Thunderstorm forecasts (Bosart and Landin 1994); CG lightning, severe weather, and mesocyclone forecasts (Jincai et al. 1992, Doswell and Flueck 1989).

When forecasts are expressed **categorically** (or are reduced from probabilistic to categorical forecasts), verification studies have typically used the Probability of Detection (POD), Probability of False Detection (POFD), False Alarm Rate (FAR), and Critical Success Index (CSI) to measure forecast quality. These measures have been used to verify: NWS temperature forecasts (Dagostaro et al. 1995); CG lightning, severe weather, and mesocyclone forecasts (Jincai et al. 1992, Doswell and Fluck 1989). The Storm Prediction Center also uses measures like these to verify their severe weather and tornado watches, as well as NWS WFO-issued weather warnings

(Crowther and Halmstad 1994, Anthony 1990, Pearson and Weiss 1979, Galway 1967). The verification procedures for SPC-issued watches and NWS WFO-issued warnings are summarized in Weiss (et al. 1980). Verification measures for NWS forecasts of temperature, probability of precipitation (PoP), precipitation type, snow amount, cloud amount, surface wind, ceiling height, and visibility are summarized in, for example, Dagostaro and Dallavalle (1991); the National Weather Service Verification Plan (National Weather Service 1985); and Dagostaro et al. (1989).

A measures-oriented approach typically fails to offer a comprehensive look at forecast performance because individual scalar measures only offer a one-dimensional view and can't identify specific strengths and weaknesses of a forecasting system (Murphy and Winkler 1987). The Brier score, for example, can identify whether a set of forecasts is more or less accurate than another, but it doesn't give any information about where in the forecast range the forecasts are more or less accurate. Indeed, this is the type of information that is lost when reducing the multi-dimensional distribution of forecasts and observations down to these one-dimensional measures.

The **distributions-oriented approach** uses the information contained in the joint distribution of forecasts and observations and so, can provide more insight into the specific strengths and weaknesses of a forecast system (Murphy and Winkler 1987). The utility of this approach has been demonstrated in the evaluation of temperature forecasts (Murphy 1989, Brooks and Doswell 1996), precipitation forecasts (Murphy et al. 1985, Murphy and Daan 1984), and tornado forecasts

(Murphy and Winkler 1982). This approach hasn't been used as extensively operationally, however. The multiple types of forecast phenomena and different spatial scales of the forecasts make this study unique.

A possible problem with this approach is that the data sets required to perform the verification can become very large. Murphy (1991) shows that the dimensionality, D, of the verification of a set of forecasts, I, and observations, K, is given by, $D = I*K - 1$. In this study, the area forecasts, for example, have 15 possible forecasts (probability values) with 2 possible observations (hit or non-event), $D = (13)(2) - 1 = 25$. Evaluating a database of this size is not prohibitive. If one wants to compare two forecast strategies, however, the dimensionality increases dramatically and is given by, $D = I*J*K - 1$, where I is the number of forecasts from the first forecast strategy, J is the number of forecasts from the second forecast strategy, and there are K observations. The amount of data needed to perform a verification of this size can become prohibitively large. If, for example, two forecast strategies are being compared for forecasting cloud intervals with 11 possible forecasts and 11 possible observations (e.g., intervals of .1 from 0 to 1), the dimensionality becomes, $D = (11)(11)(11) - 1 = 1320$ (Brooks and Doswell 1996).

In this paper, both the measures- and distributions-oriented approaches are used to evaluate the forecasts made during VORTEX '94, with the main purposes being to: 1) assess the quality of the forecasts made, 2) learn more about probabilistic forecasting and its potential application to operational forecasting, and 3) learn more

about the measures-oriented and distributions-oriented approaches to forecast verification. "Only through looking at the joint distribution of forecasts and observations, and the conditional and marginal distributions derived from the joint distribution, is it possible to get a complete picture of forecast quality" (Murphy 1993).

CHAPTER 2

THE VORTEX 94 EXPERIMENT

2.1 Purpose of the experiment

The Verification of the Origins of Rotation in Tornadoes Experiment (VORTEX) was conducted during the springs of 1994 and 1995 (01 April to 15 June) to evaluate a set of hypotheses pertaining to tornadogenesis and tornado dynamics. To evaluate these hypotheses, field operations involving storm intercept activities needed to be in the immediate vicinity of tornadic or potentially tornadic thunderstorms. Field operations also included post-intercept surveys, to be conducted on non-event days. A description of the field operations can be found in a paper by Rasmussen *et al* (1994).

An ancillary goal of VORTEX was to explore new approaches and techniques for forecasting tornadoes and tornadic storms. Forecasts made during VORTEX served both to support field operations, and to test experimental forecasting techniques, as in other recent field experiments (Doswell and Flueck 1989; Jincai et al. 1992). Not all of the experimental forecasting techniques are evaluated in this current study. The success of the field operations depended, to some extent, on the accuracy of the forecasts provided as guidance to the field coordinator (FC), who made the final deployment or non-deployment decision. One of the purposes in evaluating these forecasts is to improve our support to future field operations similar to VORTEX.

2.2 Forecasting for a field experiment

Forecasting for a field project differs from operational forecasting in many ways (Doswell et al. 1986). Project forecasters are asked for more specific forecasts, in both time and space, of a few specific weather events rather than the broad range of operational forecasts made, for example, by the National Weather Service. Besides their forecasts, the field project forecasters also may have the added responsibility of coordinating with chase teams, who need their immediate assessment of the current weather conditions in terms of a "nowcast". An added stress for field project forecasters is that the false deployment of field teams can be very expensive, putting them under considerable pressure to forecast correctly for financial reasons, not just for their personal pride and reputation. In a field project, there often are new products available for the forecasters to consider. This means that more data are available for analysis than would normally be seen in an operational setting, and the additional data may not be familiar to them. The new data can create "information overload" and add to uncertainty rather than reducing it. Thus, even previously experienced forecasters may be challenged to perform well in a field project. Forecasts in subsequent years of a field project (e.g., second year of VORTEX) should improve over the first year by virtue of experience, assuming the same forecasters are used (Doswell and Flueck 1989).

2.3 Forecasting for VORTEX 94

2.3.1 The forecasters

The forecasters who participated in this experiment (Fig. A1, Appendix A) were chosen based on their experience and skill at forecasting severe thunderstorm initiation in terms of both location and timing. It is believed that these forecasters include representatives from among the best in the field for forecasting severe storms. This prior experience was important in trying to get the field teams deployed to within striking distance of that day's convection. If the forecast target area was too far away from the actual target area, the field teams would not have time to reposition. This means a loss of valuable resources as well as missed opportunities for data collection. Owing to resource limitations, the forecasters were given no training except to familiarize themselves with the computer systems they would be using and the format of the forecasts they were to make.

For each day of the experiment (01 April to 15 June), there was a lead forecaster (LF) who was responsible for producing mesoscale forecasts for the FC and the Aircraft Coordinator, as well as an assistant forecaster (AF) who was responsible for assisting in forecast preparation. The LF and AF worked in two shifts, with LF shifts from 0700-1500 LT (day shift) and 1430-2230 LT (evening shift), and AF shifts from 0630-1430 LT (day shift) and 1400-2200 LT (evening shift). The day shift was mostly concerned with picking a preliminary target area and the issuance of a forecast,

while the evening shift monitored the afternoon and evening convection, updated forecasts, and advised field teams.

2.3.2 The forecasts

The forecasts made during VORTEX were designed to be for verifiable events within the VORTEX forecast area. That is, the necessary observational data were to be readily available for verification. There were four basic types of products issued by the forecasters: 1) VORTEX area forecasts (Day-1 and Day-2 outlooks) - designed to forecast **whether** the phenomena were going to occur, 2) VORTEX contour forecasts (Day-1 graphical probabilistic forecasts) - designed to forecast **where** storms would develop, 3) Nowcasts - providing updates to previously issued forecasts, and 4) Forecast soundings.

2.3.2.1 VORTEX area forecasts - Day-1 and Day-2 outlooks

The VORTEX area forecasts (Day-1 and Day-2 outlooks) were issued by 1400Z to forecast cloud-to-ground (CG) lightning strikes, severe weather, tornadoes, and targetable storms. Targetable storms were defined as supercellular or tornadic non-supercellular storms (e.g., landspouts). The criteria for defining a targetable storm included a storm having a confirmed mesocyclone. Finding mesocyclones in the data would require an exhaustive look at all of the WSR-88D Doppler radar data for each of the radar sites within the VORTEX area. This was deemed to be too expensive

(about $24,000 for all the tapes needed from NCDC) and too time-consuming at this point. Therefore, the definition of targetable storm was modified to use NWS/WFO-issued tornado warnings as an indication of mesocyclones. Since most tornado warnings are associated with either a mesocyclonic signature or a spotter report of a tornado (in which case, it's a targetable storm anyway), most tornado warning areas are targetable storm areas. There will be targetable storms that were missed, however, when there was a mesocyclone present on radar, but the forecaster didn't deem it necessary to issue a tornado warning. This should be a small minority of the mesocyclones, but there is no way to be certain until a comprehensive look at all of the radar data can be done.

Another criterion for defining a targetable storm was any storm that was **targeted** by the field teams. It should be obvious that not all the days with field operations resulted in successful intercepts of rotating supercells or tornadic non-supercellular storms. To be considered targetable, storms only had to be selected as a target by the field teams for mobile observations. This means that even towering cumulus could be considered a targetable storm, if so chosen by the field teams. Since there were relatively few storms targeted by the field teams (most targetable storms were from tornado or tornado warning data), this assumption will not affect the results significantly.

The Day-1 and Day-2 outlooks included: the probability of occurrence for CG lightning strikes, severe weather, tornadoes, and targetable storms, a forecast time for

the first deep convection, first severe report, and first tornado, the expected pre-supercell and supercell storm motions, and a narrative discussion (see Fig. A2 in Appendix A for a sample Day-1 outlook). The forecast probabilities were the probabilities of seeing at least one hit (event occurrence) within the VORTEX area in the appropriate time frame, which was 14-04Z (current day) for Day-1 forecasts and 11-04Z (next day) for Day-2 forecasts. The VORTEX forecast area was defined as; 31.5-39.5 degrees latitude and 94.5-103 degrees longitude (Fig. A3, Appendix A). There was **one** probability for the whole VORTEX forecast area and only **one** occurrence was needed to count a hit for the day.

2.3.2.2 VORTEX contour forecasts (Day-1 graphical probability forecasts)

The VORTEX contour forecasts were made for: CG lightning strikes within an MDR box[1], targetable storms within an MDR box, the conditional probability of tornadoes within an MDR box given that there was a tornado in the VORTEX area (hereafter called "tornado given tornado" or TGT), and the conditional probability of tornadoes within an MDR box given that there was lightning within that particular MDR box (hereafter called "tornado given lightning" or TGL). In contrast to the area forecast above, where there was one probability for the whole VORTEX forecast area, the contour forecasts assigned a probability value to each MDR box within the

[1] The standard manually-digitized radar (MDR) grid is described in section 2.6 and is shown in Fig. A4 in Appendix A. Note that the grid lines are curved on this latitude-longitude map, but they would be square boxes (nominally 47.625 km on a side or 1/4th the LFM (limitied-area fine-mesh) grid size on a polar stereographic projection.

VORTEX forecast area, with hits (and non-events) being counted separately for each probability category. The forecasters prepared hand drawn probability contours over the VORTEX domain for each of the above categories using the following contour intervals: 01, 10, 20, 40, 60, 80, 90, and 99% (see Fig. A5 in Appendix A for the sample contour forecasts). Within any given contour, probabilities were not interpolated; rather, they were assigned the value of the bounding contour. The contour intervals 1% and 99% contours were used by the forecasters to depict areas where the probability was less than 1% (event nearly certain to not occur) and greater than 99% (event nearly certain to occur). Those boxes not covered by contours, to the left of the 1% contour, were considered to have a forecast probability of zero, although zero was never explicitly contoured.

2.3.2.3 Nowcasts and forecast soundings

Nowcasts were issued as needed to update the Day-1 outlook and, since they were narrative discussions (similar to convective discussions issued by the Severe Local Storms Unit (SELS) of the National Severe Storm Forecast Center (NSSFC)), it was never intended that these would be part of the verification. They were used to either: 1) supplement or update information already contained in a valid Day-1 outlook, or 2) provide short-term details of changes in storm structure or environment after a target had been selected. The forecast soundings, prepared using SHARP software, were used as initialization input to the Advanced Regional Prediction

System (ARPS) model, and were also designed to be a follow-up to the STORMTIPE experiment conducted in 1991 (Brooks et al. 1993). The forecast soundings will not be verified in this study.

2.4 Use of forecasts in VORTEX field deployment decisions

The morning forecast was issued by 0900 LT so that decisions could be made, by the FC, concerning possible field team deployment. This early forecast time was chosen so that field teams would have time to position themselves in that day's target area before the afternoon convection began. Note that this is before the operational numerical weather prediction (NWP) model runs are available. There was also a forecast made for the following day so, if conditions weren't favorable on the current day, the FC had the option of pre-positioning the field teams closer to the next day's target area. Pre-positioning of field teams was used infrequently, and only when it was known that the next day's activity would be too far away to reach if the FC waited until the following morning to make a decision. Updates to the forecasts were made as needed to reflect the latest environmental conditions, with the hope that later forecasts would become more time- and space-specific, to pinpoint the target area as closely as possible. The forecaster's input to the FC was obviously not the only input in the decision process. Other factors play a role in this decision, including: whether or not field teams were deployed the previous day and at what time they returned, whether or not there were vehicles severely damaged by weather the previous day, the

distance to the target area and whether it could be reached during daylight hours, and how much money was left for the project.

The contour forecasts were designed to identify the "where" of the deep convection, targetable storms, and tornadoes. The intent of the targetable storm initiation forecast was to guide the FC and the associated field teams to an area where they would be most likely to find storms worthy of targeting. The tornado given tornado forecasts identified the expected location of tornadoes, if they were to occur in the VORTEX area. In other words, given that a tornado will occur somewhere in the VORTEX area, the forecasters were asked to specify the probability of its occurrence in each MDR box. The tornado given lightning forecasts gave the likelihood of seeing a tornado in an MDR box, if convection were to occur in the same MDR box. In other words, if a thunderstorm occurs in an MDR box, the forecast is to estimate the likelihood of seeing a tornado in that same box. Note that these conditional tornado probabilities are for individual MDR boxes, while the unconditional tornado probabilities are for the probability of seeing an event in the VORTEX area. Once field teams were in a general target area, this information was helpful in directing them to the storm that was more likely to become tornadic when there were multiple storms.

2.5 Observational data

Hits for each of the weather phenomena were determined as follows: **deep convection (lightning)** = detected CG lightning strikes, **severe storms** = winds in

excess of 50 kts and/or hail greater than 3/4" in diameter and/or tornadoes, **targetable storms** = supercellular storms or non-supercellular tornadic storms (tornado and tornado warning logs), and **tornadoes** = tornado report listed in the SELS log. The data available to perform the verification came from many sources. The lightning data was from the National Lightning Detection Network (NLDN) which records cloud to ground strikes (CGs) detected by a network of sensors throughout the U.S. The logs for severe weather reports, tornado reports, and tornado warnings were all obtained from the NSSFC/SELS unit in Kansas City.

The tornado log had all of the same reports found in *Storm Data* and the tornado warning log had the same information we had already obtained from the NWS/WFOs, but the severe log had some "missing" reports (249 reports missing in the sense that they were listed in *Storm Data* but not listed in the SELS log of severe reports). This is because the SELS log discards reports that are within 10 miles and 15 minutes of a previous report. These were added back in to make the log more complete but, it made little difference in the overall result because virtually all of these added reports were in the same grid boxes that had already received previous reports.

2.6 Gridding forecasts and observations on an MDR grid

Verification of the contour forecasts required that forecast probabilities be assigned to each grid box as well as assigning hits (or non-events) to each grid box according to the observed weather phenomena. The forecasts were gridded by

recreating the hand drawn contours on the PC-McIDAS computer system using a probability contouring program that assigned probabilities to the MDR boxes. The observed events were obtained from the sources listed earlier with locations given in latitude-longitude coordinates. These latitude-longitudes were converted to (x, y) coordinates on the MDR grid using the following equations for the polar stereographic projection:

$$x = 6370(\sigma \cos\phi \cos\lambda) \tag{1}$$

$$y = 6370(\sigma \cos\phi \sin\lambda) \tag{2}$$

$$\sigma = \left(\frac{1 + \sin\left(\frac{\pi}{3}\right)}{1 + \sin\phi} \right) \tag{3}$$

where ϕ is the latitude, λ is the deviation of longitude from the standard longitude (105° W), and σ is the image scale factor. Using these (x, y) coordinates, the event is assigned to a particular MDR box by comparing the event coordinates with the MDR box coordinates. One or more occurrences in a given MDR box was considered a hit for that box. The hits (and non-events) were then summed for each forecast probability category. For example, all hits (and non-events) that occurred in MDR

boxes with 20% forecast, were assigned to the 20% category in the contingency table. (see section 4.2.2 and Tables 2i-l).

CHAPTER 3

THE JOINT DISTRIBUTION APPROACH

3.1 Introduction

A distributions-oriented approach to forecast verification (Murphy and Winkler 1987) looks at the distribution of forecasts and observations and attempts to explain the relationships between them. Our forecasts are probability values and our observations are either hits ($x = 1$) or non-events ($x = 0$) for the different phenomena that we are attempting to forecast. Ideally, high relative frequencies would be assigned to (f, x) pairs where f is equal to or close to x, and low relative frequencies to (f, x) pairs where f is not close to x. A perfect forecast is realized when all relative frequencies are equal to zero except for the pairs (f=1, x=1) and (f=0, x=0). There are two factorizations of this joint distribution that will be explored in the discussion that follows.

3.2 Factorization of the joint distribution

3.2.1 Calibration-refinement factorization $p(f, x) = p(x \mid f)\, p(f)$

This factorization explores the conditional distribution of the forecasts given the observations $p(x \mid f)$ (the **calibration**) and the marginal distribution of the

forecasts $p(f)$ (the **refinement**). For a perfectly reliable (or perfectly calibrated) forecast, the observed relative frequency equals the forecast probability. In other words, for good reliability, hits should be observed 40% of the time that 40% is forecast. The marginal distribution $p(f)$ tells how often different forecast values are used. If the same probabilities are used on event days as on non-event days (e.g., a forecast of climatology), the forecasts are not refined though they would be perfectly reliable. On the other hand, forecasts that use a probability of 100% on event days and 0% on non-event days, are perfectly refined and perfectly calibrated. A refined forecast doesn't necessarily imply a well-calibrated forecast, however. Suppose a forecaster attempts to show perfect refinement by only using categorical forecasts of one and zero, with the $p(f = 1) = .4$ and $p(f = 0) = .6$, but that $p(x = 1|f = 1) = p(x = 1|f = 0) = .4$. An event then, is equally likely, regardless of the forecast. This is a forecast that appears refined, but isn't at all calibrated. The goal, then, is to have forecasts that are both well-calibrated and refined (Murphy and Winkler 1987).

A related quality indicator using the conditional probability, $p(x|f)$, is the resolution. The resolution tells how well the forecasts distinguish between days that are more or less likely than climatology to observe an event. Reliable forecasts don't always show good resolution. Constant forecasts of climatology, for example, would show no resolution though they would have perfect reliability. Better resolution is obtained when the observed relative frequencies are much different from climatology.

A forecast for which the $p(x \mid f) = p(x)$ shows no resolution since the observations are independent of the forecasts.

3.2.2 Likelihood-base rate factorization $p(f, x) = p(f \mid x)\, p(x)$

This factorization looks at the conditional distribution of the forecasts given the observations $p(f \mid x)$ (the **likelihood**), and the marginal distribution of the observations $p(x)$ (the **base rate**). That is, given an observation, what is the likelihood, $p(f \mid x)$, that it was forecast correctly and what is the base rate (or sample climatology) of the events, $p(x)$? Most verification studies focus on the distribution of the observations after a set of forecasts was made, whereas a likelihood-base rate factorization looks at the distribution of the forecasts, given a set of observations. The value in doing so is contained within the understanding of the overall relationships between forecasts and observations (Murphy and Winkler 1987). The conditional probability $p(f \mid x)$ can tell us how well the forecasts discriminate among different observations. A forecast where the conditional probability, $p(f \mid x)$, is the same for all x does not discriminate between hits and non-events. The conditional probability, $p(f \mid x=1)$, should increase as the value of f increases, since it is desirable to have higher probabilities forecast on those occasions where a hit was observed. Conversely, the conditional probability, $p(f \mid x=0)$, should decrease as the value of f

increases, since lower probabilities should have been used on those forecast occasions where a non-event was observed. The marginal distribution, $p(x)$ (the base rate or sample climatology), is the distribution of the observations, which gives us an idea of the uncertainty in the forecast induced by the variability of the observations. When the events are highly variable, the forecasts have more uncertainty also. Forecasts in which $p(f \mid x) = p(f)$ show no discrimination since the forecasts are independent of the observations. The goal, then, is to have forecasts that are perfectly discriminatory, as well as having a representative sample climatology[2] (Murphy and Winkler 1987).

3.2.3 Relationship between factorizations

Murphy and Winkler (1987) make it clear that these two factorizations represent complementary rather than alternative ways to approach the verification problem. When evaluated together, they give a more comprehensive look at how "good" the forecasts are. It is important that there be consistency between them. A forecast that is perfectly calibrated and perfectly refined will also be perfectly discriminatory. On the other hand, a perfectly discriminatory forecast won't necessarily be well calibrated or refined. For example, if a non-event follows all forecasts of 40% and a hit always follows a forecast of 60%, the forecasts are perfectly discriminatory but aren't well-calibrated or refined.

[2] The long-term climatological frequency of these phenomena over the temporal and spatial domains used in this experiment was not available, so the sample climatologies were used.

CHAPTER 4

VERIFICATION OF VORTEX-94 FORECASTS

4.1 Introduction

The two different approaches to forecast verification are now applied directly to those forecasts made during VORTEX-94. The quality of the forecasts will be examined using the traditional measures-oriented approach first, followed by the distributions-oriented approach.

4.2 Distribution of forecasts and observations

4.2.1 VORTEX area forecasts

The probabilities forecast for each day and category are shown in Tables 1a,b. The probabilities of CG lightning strikes (L), severe weather (S), tornadoes (T), and targetable storms (TS) are presented as percentage values. These tables also show whether the events verified or not with a "Y" for yes or a "N" for no for each day and category. Verification of the Day-2 forecasts begins on 02 April, since forecasting started on 01 April. The contents of Tables 1a,b are presented as contingency tables in Tables 2a-h, where each probability category is shown against the number of hits, non-events, and forecasts for that category. Note that there are 76 possible days (01 April to 15 June) for Day-1 forecasts and only 71 possible days for the Day-2 forecasts.

Table 1a. VORTEX area Day-1 forecasts with (Y) for hit and (N) for non-event days.

Date	L	S	T	TS		Date	L	S	T	TS
4/1	80/Y	60/N	20/N	30/N		5/9	100/Y	90/Y	30/Y	50/Y
4/2	40/Y	40/Y	20/Y	20/Y		5/10	100/Y	90/Y	20/N	40/Y
4/3	02/N	02/N	00/N	00/N		5/11	100/Y	80/Y	30/Y	30/Y
4/4	80/Y	50/N	30/N	40/N		5/12	100/Y	60/Y	10/Y	20/Y
4/5	100/Y	40/Y	05/N	10/N		5/13	100/Y	70/Y	30/N	40/Y
4/6	00/N	00/N	00/N	00/N		5/14	100/Y	50/Y	30/Y	30/Y
4/7	10/Y	05/N	05/N	05/N		5/15	10/Y	05/N	02/N	02/N
4/8	40/Y	20/Y	10/N	20/N		5/16	20/Y	05/N	02/N	02/N
4/9	98/Y	95/Y	70/Y	80/Y		5/17	40/N	30/N	20/N	20/N
4/10	95/Y	80/Y	60/Y	50/Y		5/18	40/Y	10/N	05/N	05/N
4/11	100/Y	80/Y	30/N	20/N		5/19	60/Y	10/N	05/N	05/N
4/12	02/N	00/N	00/N	00/N		5/20	70/N	50/N	30/N	20/N
4/13	05/N	00/N	00/N	00/N		5/21	50/Y	40/Y	05/N	05/N
4/14	80/Y	70/Y	50/N	60/N		5/22	50/Y	10/Y	02/N	02/N
4/15	30/Y	05/Y	02/Y	05/Y		5/23	60/Y	30/Y	02/N	02/N
4/16	00/N	00/N	00/N	00/N		5/24	100/Y	40/Y	10/N	20/Y
4/17	10/N	05/N	00/N	00/N		5/25	100/Y	90/Y	50/Y	70/Y
4/18	10/N	02/N	00/N	00/N		5/26	100/Y	90/Y	30/Y	40/Y
4/19	05/N	00/N	00/N	00/N		5/27	80/Y	50/Y	10/Y	10/Y
4/20	70/Y	10/Y	02/N	00/N		5/28	50/Y	40/Y	05/Y	05/Y
4/21	100/Y	10/Y	02/N	00/N		5/29	98/Y	90/Y	70/Y	80/Y
4/22	100/Y	30/Y	02/Y	00/Y		5/30	50/N	40/N	05/N	05/N
4/23	50/Y	20/N	02/N	02/N		5/31	70/Y	50/Y	10/N	10/Y
4/24	95/Y	50/Y	20/N	30/N		6/1	80/Y	70/Y	50/N	60/N
4/25	100/Y	100/Y	95/Y	95/Y		6/2	95/Y	60/Y	10/N	10/N
4/26	70/Y	60/Y	20/Y	10/Y		6/3	100/Y	60/N	05/N	02/N
4/27	80/Y	70/Y	60/N	60/Y		6/4	95/Y	30/Y	02/N	10/N
4/28	90/Y	10/N	02/N	02/N		6/5	95/Y	60/Y	05/Y	10/Y
4/29	100/Y	100/Y	30/Y	60/Y		6/6	100/Y	50/Y	10/Y	10/Y
4/30	05/N	00/N	00/N	00/N		6/7	100/Y	95/Y	50/N	80/N
5/1	20/Y	05/N	02/N	02/N		6/8	100/Y	70/Y	30/Y	30/Y
5/2	100/Y	50/Y	20/N	30/N		6/9	100/Y	98/Y	40/Y	50/Y
5/3	95/Y	50/Y	10/Y	30/Y		6/10	100/Y	100/Y	70/Y	80/Y
5/4	90/Y	30/N	10/N	20/N		6/11	100/Y	100/Y	60/Y	80/Y
5/5	95/Y	80/Y	60/Y	80/Y		6/12	50/Y	30/Y	02/N	05/N
5/6	100/Y	95/Y	30/Y	70/Y		6/13	50/Y	30/N	02/N	05/N
5/7	100/Y	40/Y	10/N	05/N		6/14	20/Y	10/N	00/N	02/N
5/8	70/Y	50/N	20/N	10/N		6/15	50/Y	10/Y	02/Y	05/Y

Table 1b. VORTEX area Day-2 forecasts with (Y) for hit and (N) for non-events days.

Date	L	S	T	TS		Date	L	S	T	TS
4/2	90/Y	80/Y	60/Y	70/Y		5/10	100/Y	90/Y	50/N	70/Y
4/3	10/N	05/N	02/N	02/N		5/11	100/Y	80/Y	20/Y	30/Y
4/4	30/Y	30/N	20/N	30/N		5/12	100/Y	80/Y	20/Y	20/Y
4/5	98/Y	40/Y	10/N	20/N		5/13	100/Y	70/Y	30/N	30/Y
4/6	02/N	00/N	00/N	00/N		5/14	90/Y	70/Y	20/Y	20/Y
4/7	10/Y	05/N	00/N	02/N		5/15	10/Y	02/N	02/N	02/N
4/8	30/Y	10/Y	05/N	10/N		5/16	05/Y	02/N	00/N	00/N
4/9	98/Y	90/Y	70/Y	80/Y		5/17	40/N	20/N	10/N	10/N
4/10	95/Y	90/Y	60/Y	70/Y		5/18	60/Y	40/N	30/N	30/N
4/11	100/Y	70/Y	40/N	30/N		5/19	40/Y	10/N	05/N	05/N
4/12	10/N	02/N	02/N	00/N		5/20	80/N	60/N	40/N	20/N
4/13	05/N	02/N	02/N	05/N		5/21	80/Y	50/Y	20/N	20/N
4/14	20/Y	05/Y	02/N	02/N		5/22	30/Y	10/Y	02/N	02/N
4/15	20/Y	05/Y	02/Y	02/Y		5/23	50/Y	20/Y	02/N	02/N
4/16	02/N	00/N	00/N	00/N		5/24	90/Y	50/Y	05/N	05/Y
4/17	00/N	00/N	00/N	00/N		5/25	80/Y	40/Y	20/Y	20/Y
4/18	20/N	10/N	00/N	00/N		5/26	90/Y	80/Y	50/Y	50/Y
4/19	20/N	05/N	00/N	00/N		5/27	90/Y	50/Y	10/Y	20/Y
4/20	no	forecast				5/28	30/Y	30/Y	20/Y	30/Y
4/21	no	forecast				5/29	80/Y	60/Y	20/Y	20/Y
4/22	100/Y	10/Y	02/Y	00/Y		5/30	60/N	50/N	20/N	20/N
4/23	no	forecast				5/31	60/Y	50/Y	40/N	40/Y
4/24	70/Y	50/Y	20/N	30/N		6/1	90/Y	70/Y	40/N	40/N
4/25	95/Y	70/Y	20/Y	30/Y		6/2	90/Y	50/Y	10/N	10/N
4/26	no	forecast				6/3	90/Y	30/N	05/N	05/N
4/27	50/Y	40/Y	30/N	30/Y		6/4	95/Y	30/Y	10/N	10/N
4/28	50/Y	40/N	10/N	10/N		6/5	50/Y	40/Y	10/Y	20/Y
4/29	100/Y	80/Y	40/Y	50/Y		6/6	50/Y	30/Y	05/Y	05/Y
4/30	80/Y	20/N	10/N	10/N		6/7	70/Y	50/Y	20/N	30/N
5/1	20/Y	10/N	02/N	02/N		6/8	98/Y	90/Y	40/Y	60/Y
5/2	70/Y	30/Y	10/N	10/N		6/9	90/Y	40/Y	10/Y	20/Y
5/3	50/Y	30/Y	05/Y	05/Y		6/10	100/Y	90/Y	30/Y	50/Y
5/4	90/Y	40/N	05/N	20/N		6/11	100/Y	95/Y	70/Y	80/Y
5/5	95/Y	50/Y	30/Y	40/Y		6/12	80/Y	60/Y	30/N	40/N
5/6	98/Y	90/Y	40/Y	80/Y		6/13	70/Y	50/N	05/N	20/N
5/7	90/Y	40/Y	05/N	10/N		6/14	50/Y	30/N	02/N	05/N
5/8	95/Y	70/N	40/N	30/N		6/15	30/Y	20/Y	02/Y	05/Y
5/9	98/Y	70/Y	30/Y	30/Y						

Table 2a. Hits (H) and non-events (N) by probability category for lightning Day-1.

Prob(%)	H	N	# Forecasts
0	0	2	2
2	0	2	2
5	0	3	3
10	2	2	4
20	3	0	3
30	1	0	1
40	3	1	4
50	7	1	8
60	2	0	2
70	4	1	5
80	6	0	6
90	2	0	2
95	7	0	7
98	2	0	2
100	25	0	25
Totals	64	12	76

Table 2b. Hits (H) and non-events (N) by probability category for severe Day-1.

Prob(%)	H	N	# Forecasts
0	0	6	6
2	0	2	2
5	1	5	6
10	4	4	8
20	1	1	2
30	4	3	7
40	6	1	7
50	7	3	10
60	4	2	6
70	5	0	5
80	4	0	4
90	5	0	5
95	3	0	3
98	1	0	1
100	4	0	4
Totals	49	27	76

Table 2c. As in Table 2a except for tornado forecasts.

Prob(%)	H	N	# Forecasts
0	0	10	10
2	3	12	15
5	2	6	8
10	4	6	10
20	2	7	9
30	7	4	11
40	1	0	1
50	1	3	4
60	3	1	4
70	3	0	3
80	0	0	0
90	0	0	0
95	1	0	1
98	0	0	0
100	0	0	0
Totals	27	49	76

Table 2d. As in Table 2a except for targetable storm forecasts.

Prob(%)	H	N	# Forecasts
0	1	11	12
2	0	9	9
5	3	7	10
10	5	4	9
20	3	5	8
30	4	4	8
40	3	1	4
50	3	0	3
60	2	2	4
70	2	0	2
80	5	1	6
90	0	0	0
95	1	0	1
98	0	0	0
100	0	0	0
Totals	32	44	76

Table 2e. As in Table 2a except for lightning Day-2 forecasts.

Prob(%)	H	N	# Forecasts
0	0	1	1
2	0	2	2
5	1	1	2
10	2	2	4
20	3	2	5
30	5	0	5
40	1	1	2
50	7	0	7
60	2	1	3
70	4	0	4
80	5	1	6
90	11	0	11
95	5	0	5
98	5	0	5
100	9	0	9
Totals	60	11	71

Table 2f. As in Table 2a except for severe Day-2 forecasts.

Prob(%)	H	N	# Forecasts
0	0	3	3
2	0	4	4
5	2	3	5
10	3	3	6
20	2	2	4
30	5	3	8
40	6	3	9
50	8	2	10
60	2	1	3
70	6	1	7
80	5	0	5
90	6	0	6
95	1	0	1
98	0	0	0
100	0	0	0
Totals	46	25	71

Table 2g. As in Table 2a except for tornado Day-2 forecasts.

Prob(%)	H	N	# Forecasts
0	0	7	7
2	3	9	12
5	2	7	9
10	3	7	10
20	7	5	12
30	3	4	7
40	3	5	8
50	1	1	2
60	2	0	2
70	2	0	2
80	0	0	0
90	0	0	0
95	0	0	0
98	0	0	0
100	0	0	0
Totals	26	45	71

Table 2h. As in Table 2a except for targetable storm Day-2 forecast.

Prob(%)	H	N	# Forecasts
0	1	7	8
2	1	7	8
5	4	4	8
10	0	8	8
20	7	6	13
30	6	6	12
40	2	2	4
50	3	0	3
60	1	0	1
70	3	0	3
80	3	0	3
90	0	0	0
95	0	0	0
98	0	0	0
100	0	0	0
Totals	31	40	71

This discrepancy is because on four of the days, forecasts weren't issued on time, and the fifth day is 01 April. Three of the four missing forecasts were issued at a later time, but these values weren't used since it would be unfair to compare these to the forecasts that were issued on time. There definitely is a trend for the forecasts to shift toward lower probabilities as the events become rarer and as the forecast range becomes longer (e.g. Day-1 to Day-2 forecasts), as expected. This is even easier to see when looking at these numbers graphically (Figs. 1a-d and 2a-d).

4.2.2 VORTEX contour forecasts

The contingency tables for the contour forecasts are shown in Tables 2i-l. There are (421 grid boxes) x (76 days) = 31,996 MDR boxes possible for the VORTEX period for CG lightning and targetable storm forecasts. There are fewer MDR boxes possible for the TGT forecasts because only 27 days observed tornadoes for (421 gridboxes) x (27 days) = 11,367 MDR boxes possible. For TGL forecasts, the total number of grid boxes possible is just those that had lightning strikes on the 64 days with observed lightning (6,562 MDR boxes).

On average, about one-fourth of the VORTEX area was covered by lightning on those days that observed lightning, whereas tornadoes and targetable storms averaged only 3 and 7 boxes per day, respectively. The extreme rarity of these phenomena is also evident looking at the events and non-events columns in Tables 2i-l and the graphs in Figs. 3a-h, where non-events vastly outnumber events.

Table 2i. As in Table 2a except for lightning contour forecasts.

Prob(%)	H	N	# Forecasts
0	407	9042	9449
1	264	4730	4994
10	563	3387	3950
20	1019	3134	4153
40	1169	2195	3364
60	1091	1694	2785
80	551	646	1197
90	838	432	1270
99	660	174	834
Totals	6562	25434	31996

Table 2j. As in Table 2a except for targetable storm contour forecasts

Prob(%)	H	N	# Forecasts
0	23	21330	21353
1	28	5358	5386
10	26	2478	2504
20	48	1549	1597
40	49	754	803
60	23	276	299
80	13	41	54
90	0	0	0
99	0	0	0
Totals	210	31786	31996

Table 2k. As in Table 2a except for the conditional probability of a tornado in an MDR box given a tornado in the VORTEX area contour forecasts

Prob(%)	H	N	# Forecasts
0	12	4861	4873
1	7	2565	2572
10	9	1596	1605
20	16	978	994
40	26	764	790
60	7	269	276
80	7	198	205
90	10	42	52
99	0	0	0
Totals	94	11273	11367

Table 2l. As in Table 2a except for the conditional probability of a tornado in an MDR box given lighning within that same MDR box contour forecasts.

Prob(%)	H	N	# Forecasts
0	16	2235	2251
1	8	2005	2013
10	16	1068	1084
20	20	689	709
40	24	443	467
60	10	28	38
80	0	0	0
90	0	0	0
99	0	0	0
Totals	94	6468	6562

Table 2m. As in Table 2i.

Prob(%)	H	N	# Forecasts
0	407	9042	9449
1	264	4730	4994
10	563	3387	3950
20	1019	3134	4153
40	1169	2195	3364
60	1091	1694	2785
80	551	646	1197
90	838	432	1270
99	660	174	834
Totals	6562	25434	31996

Table 2n. Re-binned hits and non-events by forecast probability category for targetable storm contour forecasts.

Prob(%)	H	N	# Forecasts
0	23	21330	21353
.055	54	7836	7890
.2	48	1549	1597
.4	49	754	803
.7	36	317	353
Totals	210	31786	31996

Table 2o. As in Table 2n, except for tornado given tornado contour forecasts.

Prob(%)	H	N	# Forecasts
0	12	4861	4873
.055	16	4161	4177
.2	16	978	994
.4	26	764	790
.767	24	509	533
Totals	94	11273	11367

Table 2p. As in Table 2n, except for tornado given lightning contour forecasts.

Prob(%)	H	N	# Forecasts
.005	24	4240	4264
.15	36	1757	1793
.5	34	471	505
Totals	94	6468	6562

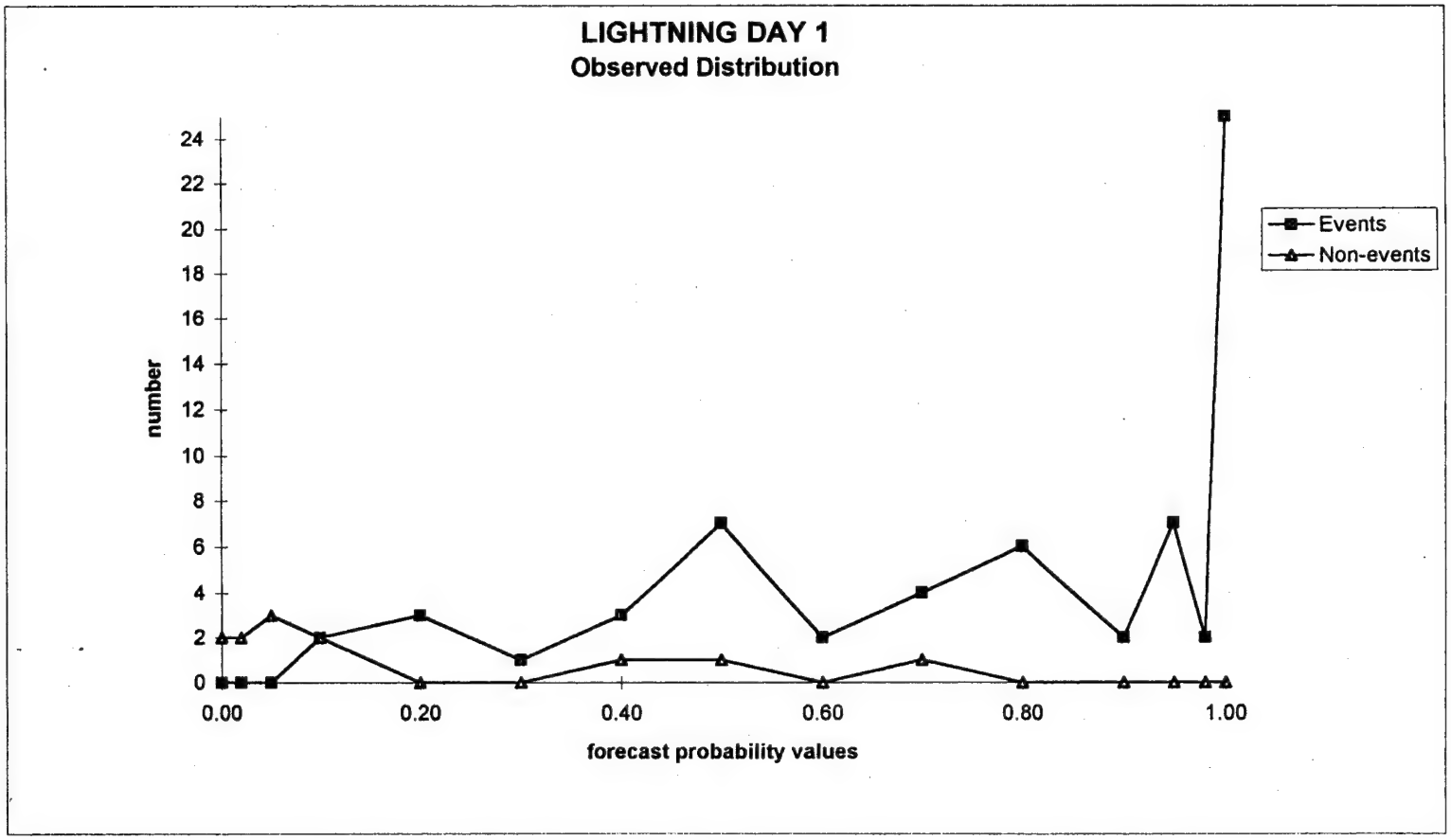


Figure 1a. Observed distributions of events and non-events by forecast probability category for lightning Day-1 forecasts.

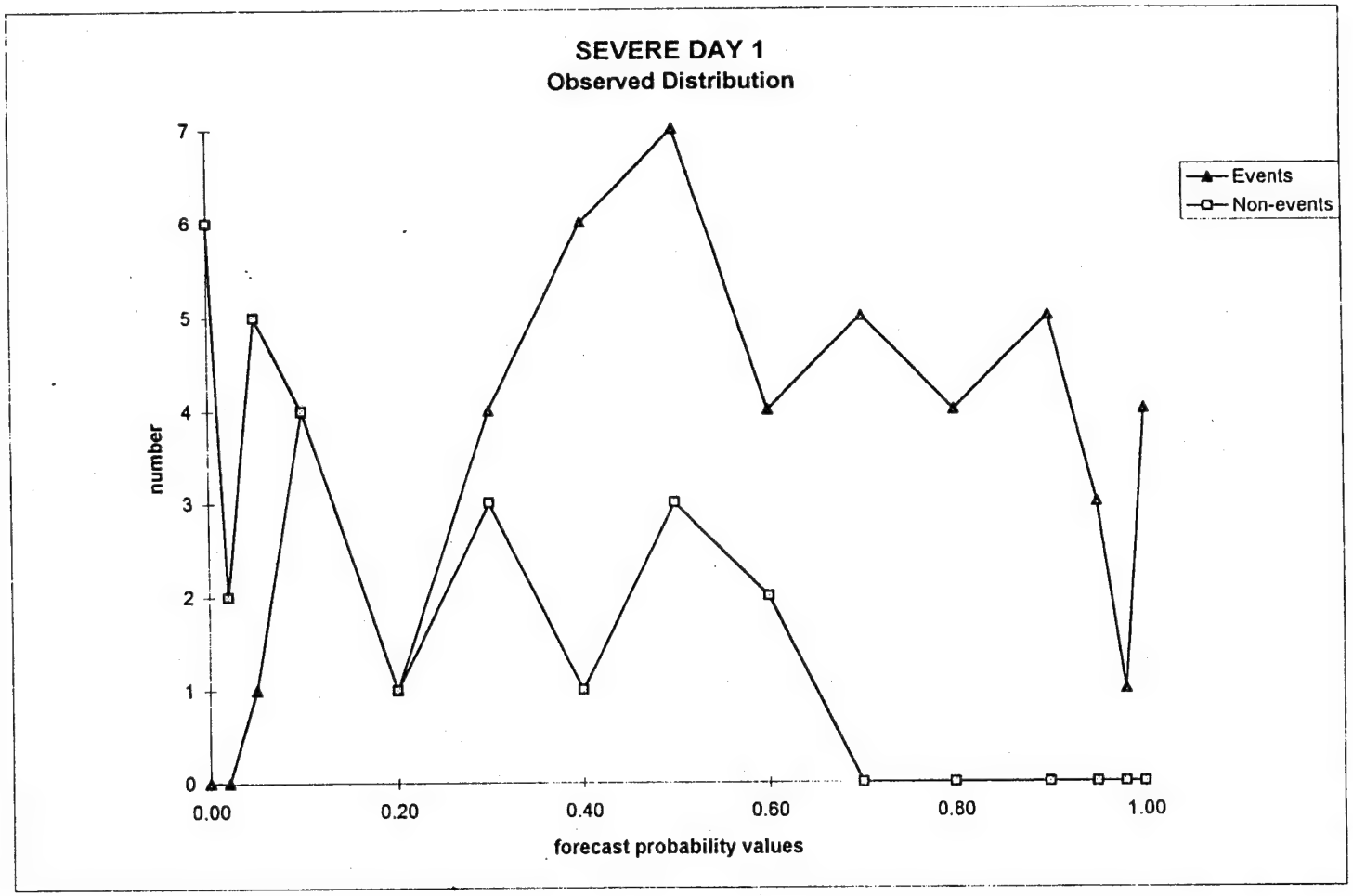


Figure 1b. As in Fig. 1a, except for severe Day-1 forecasts.

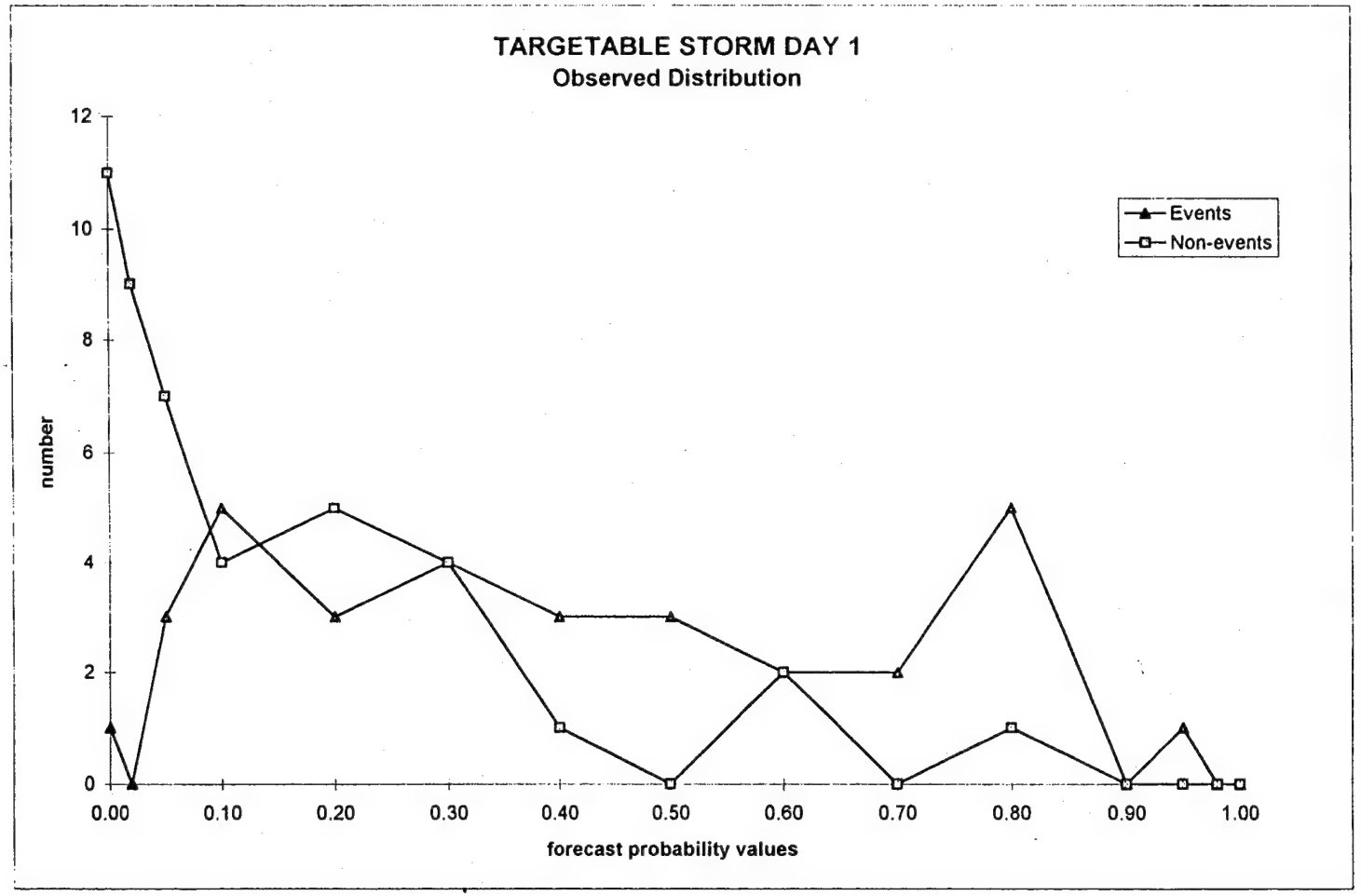


Figure 1c. As in Fig. 1a, except for targetable storm Day-1 forecasts.

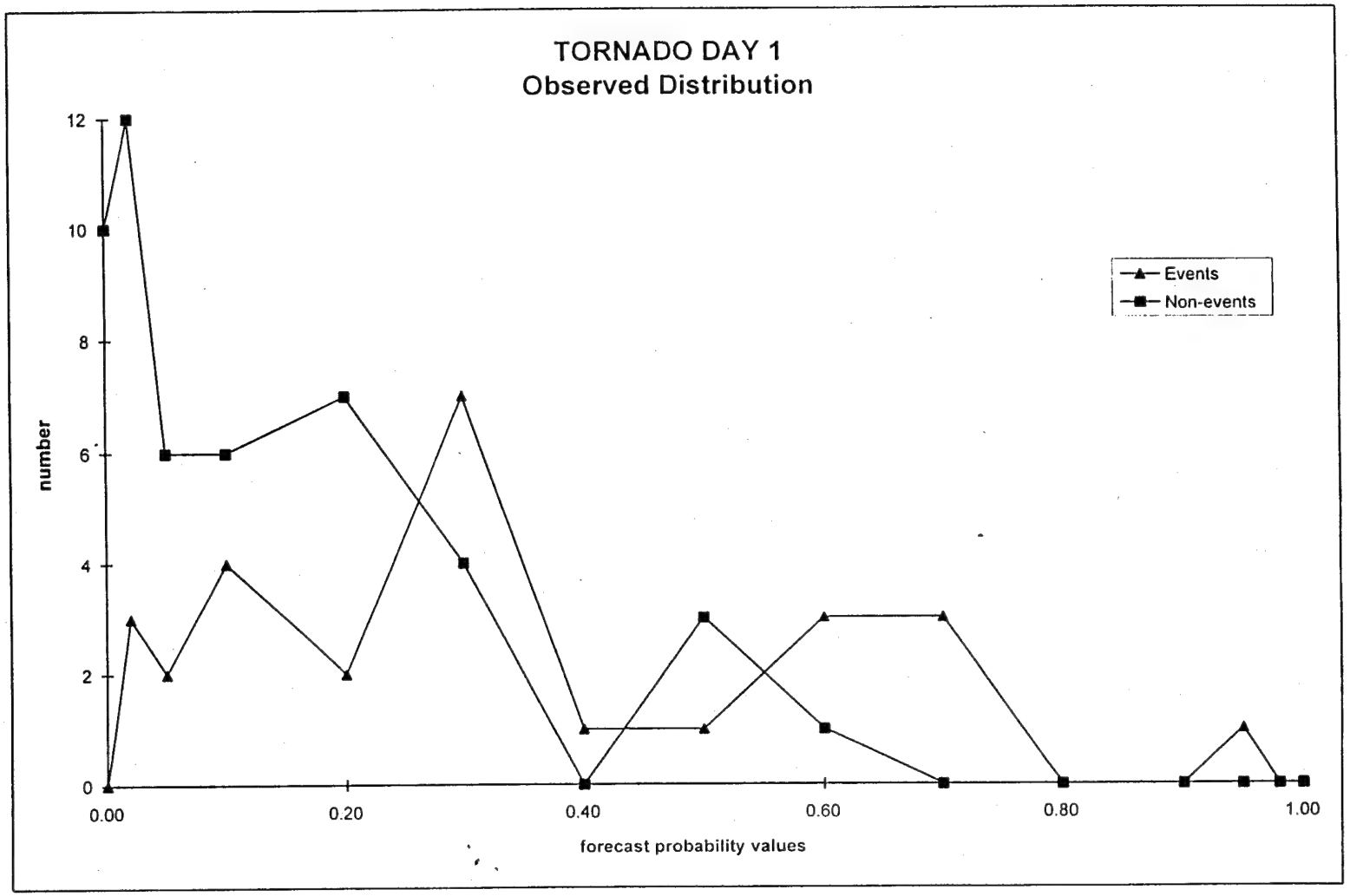


Figure 1d. As in Fig. 1a, except for tornado Day-1 forecasts.

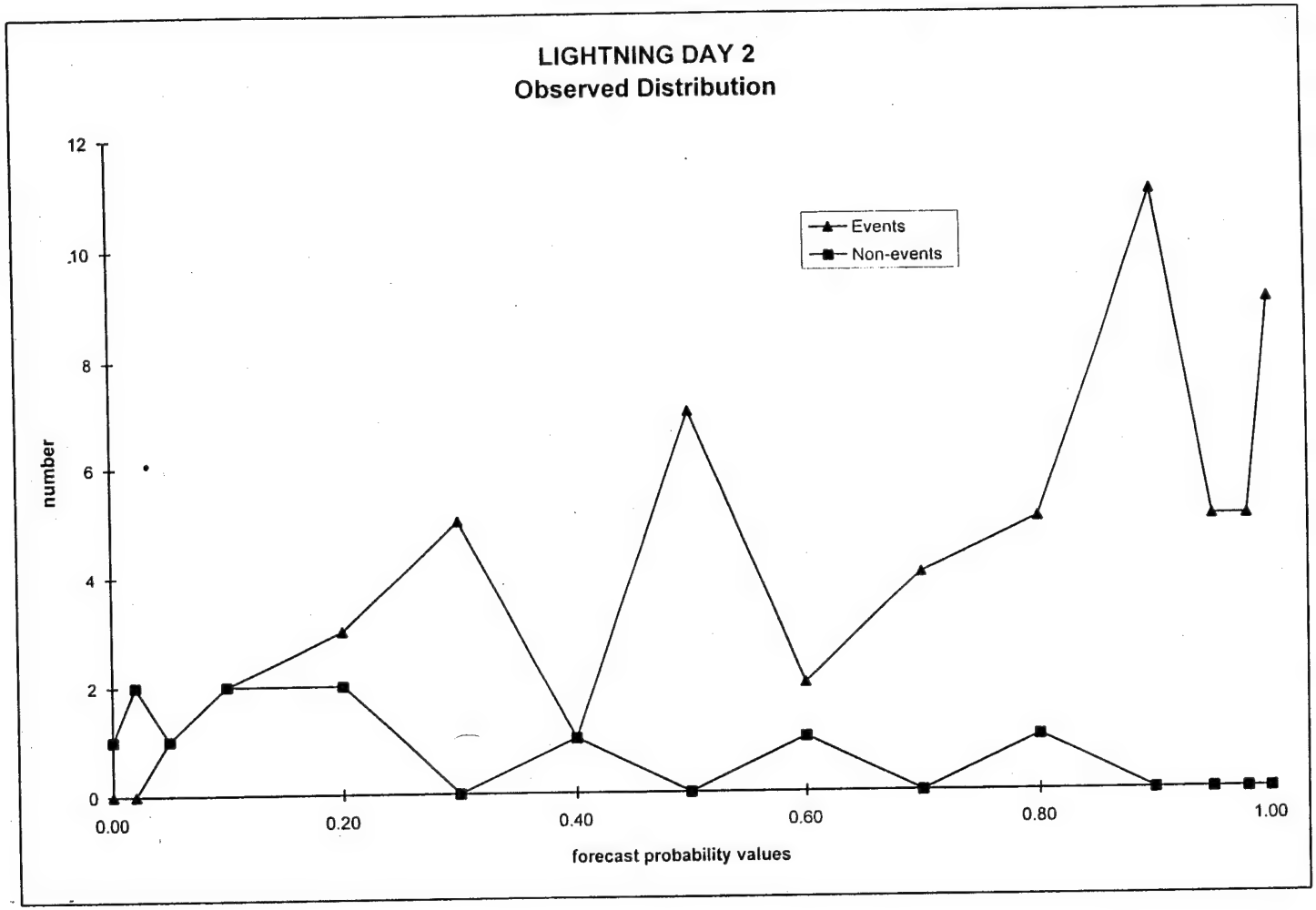


Figure 2a. Observed distribution of events and non-events by forecast probability category for lightning Day-2 forecasts.

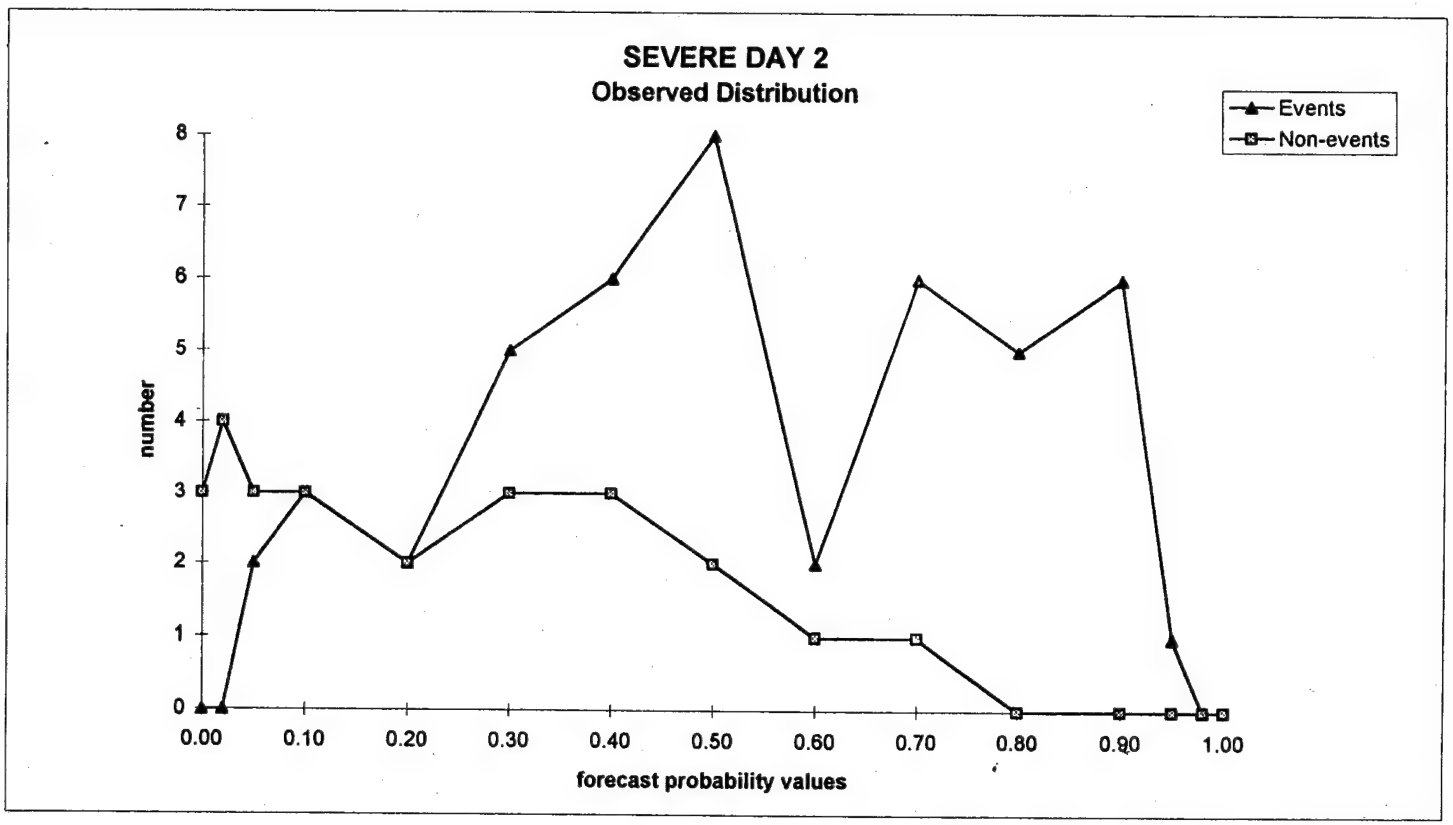


Figure 2b. As in Fig. 2a, except for severe Day-2 forecasts.

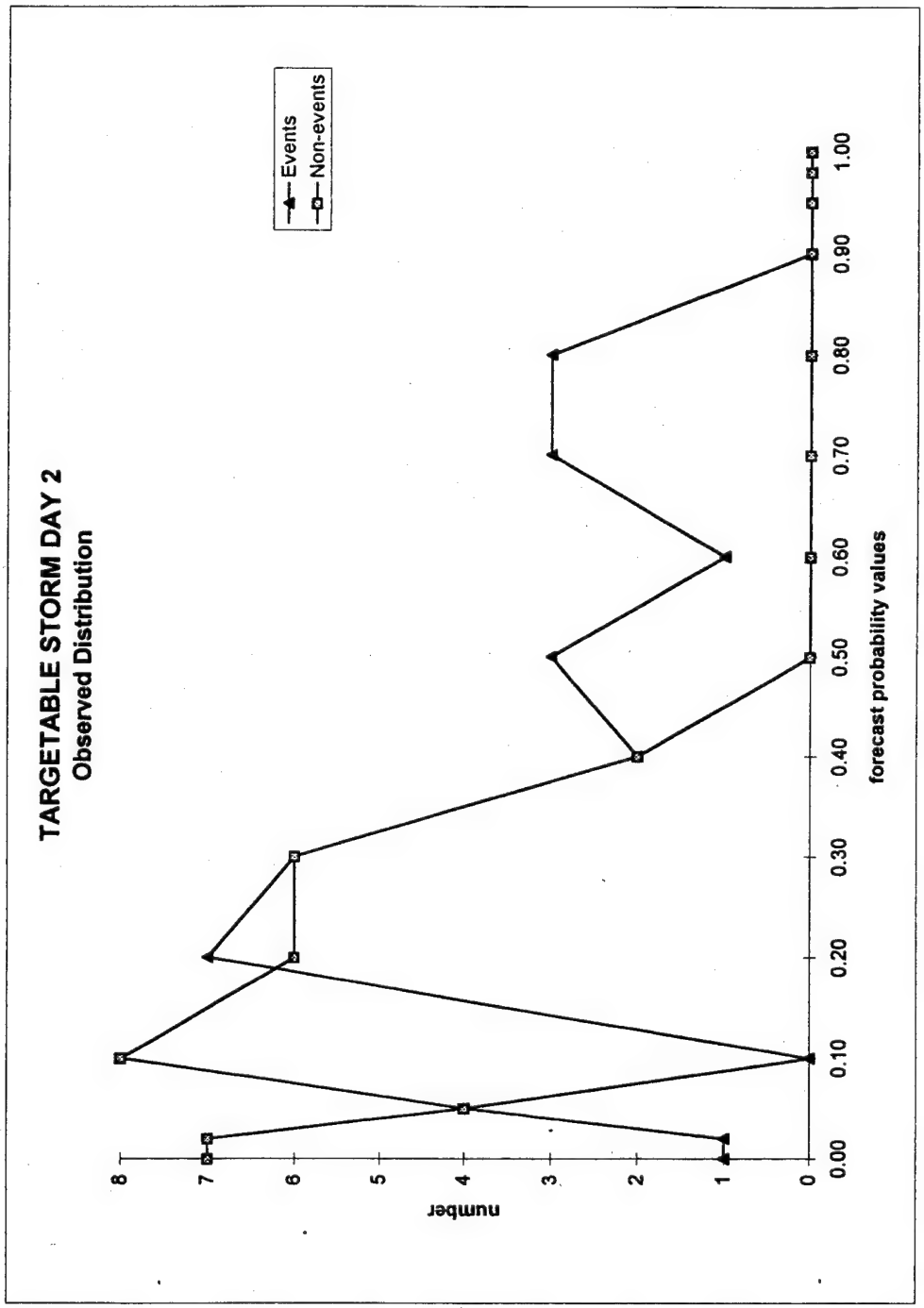


Figure 2c. As in Fig. 2a, except for targetable storm Day-2 forecasts.

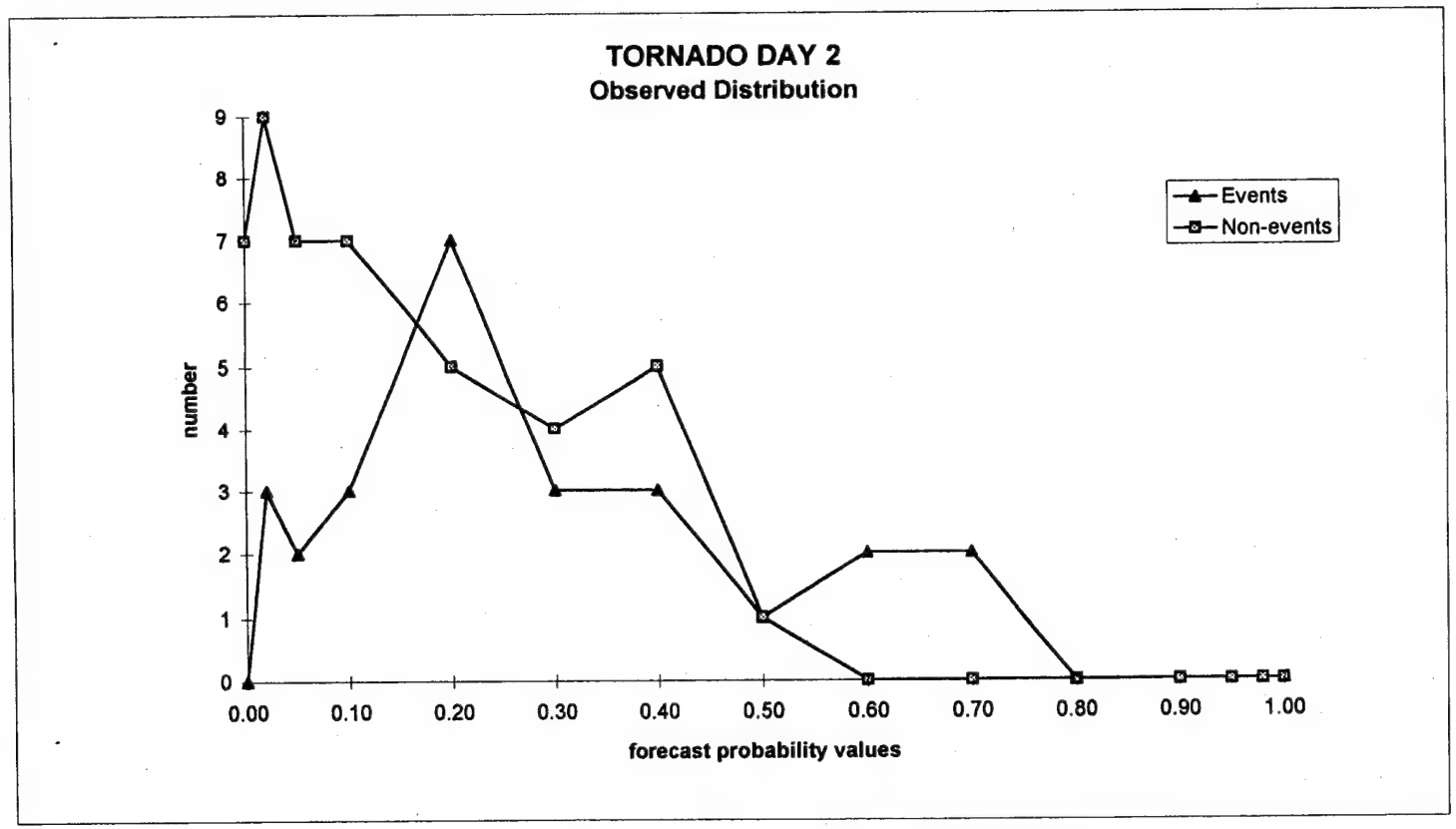


Figure 2d. As in Fig. 2a, except for tornado Day-2 forecasts.

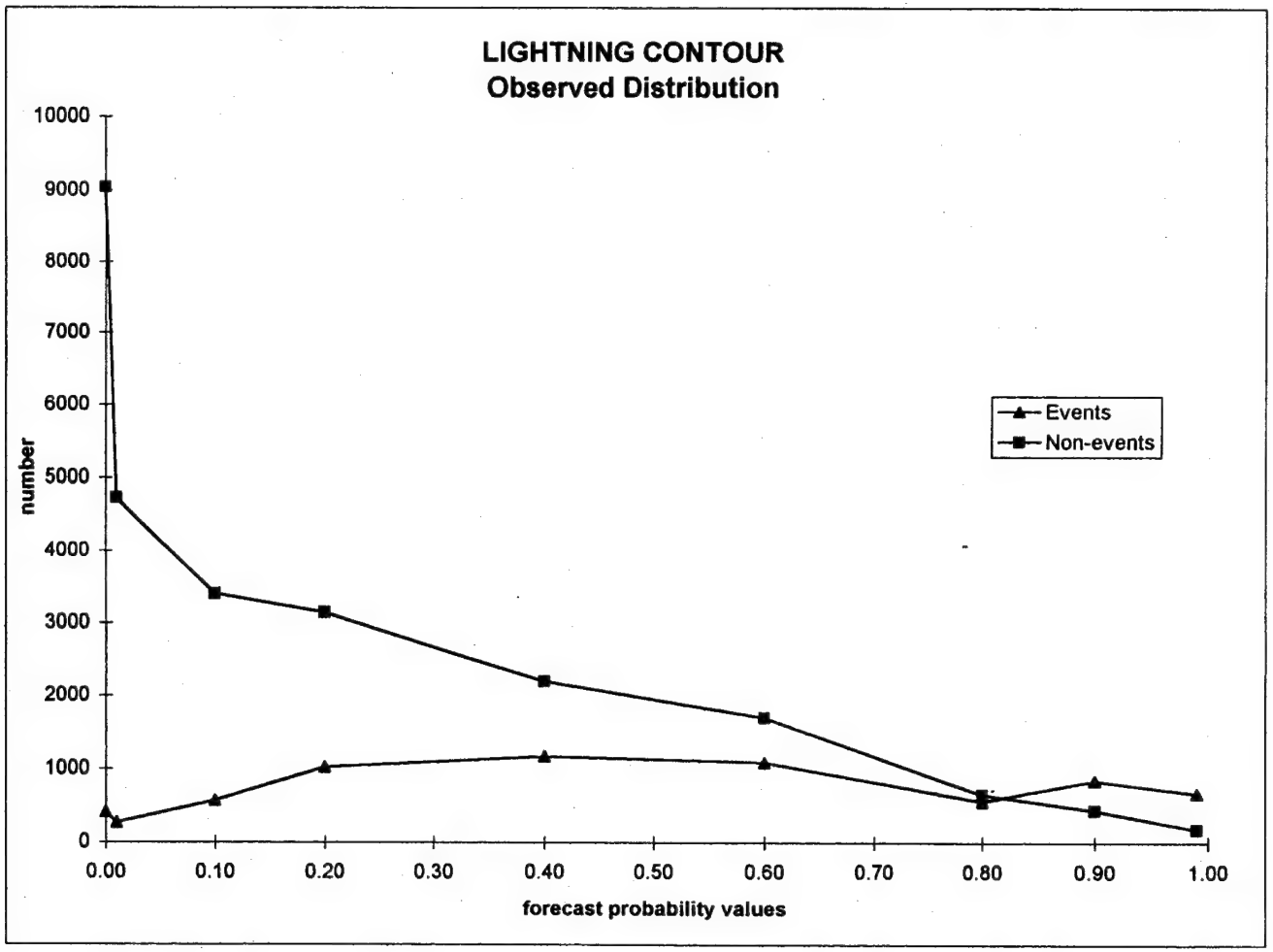


Figure 3a. Observed distribution of events and non-events by forecast probability category for lightning contour forecasts.

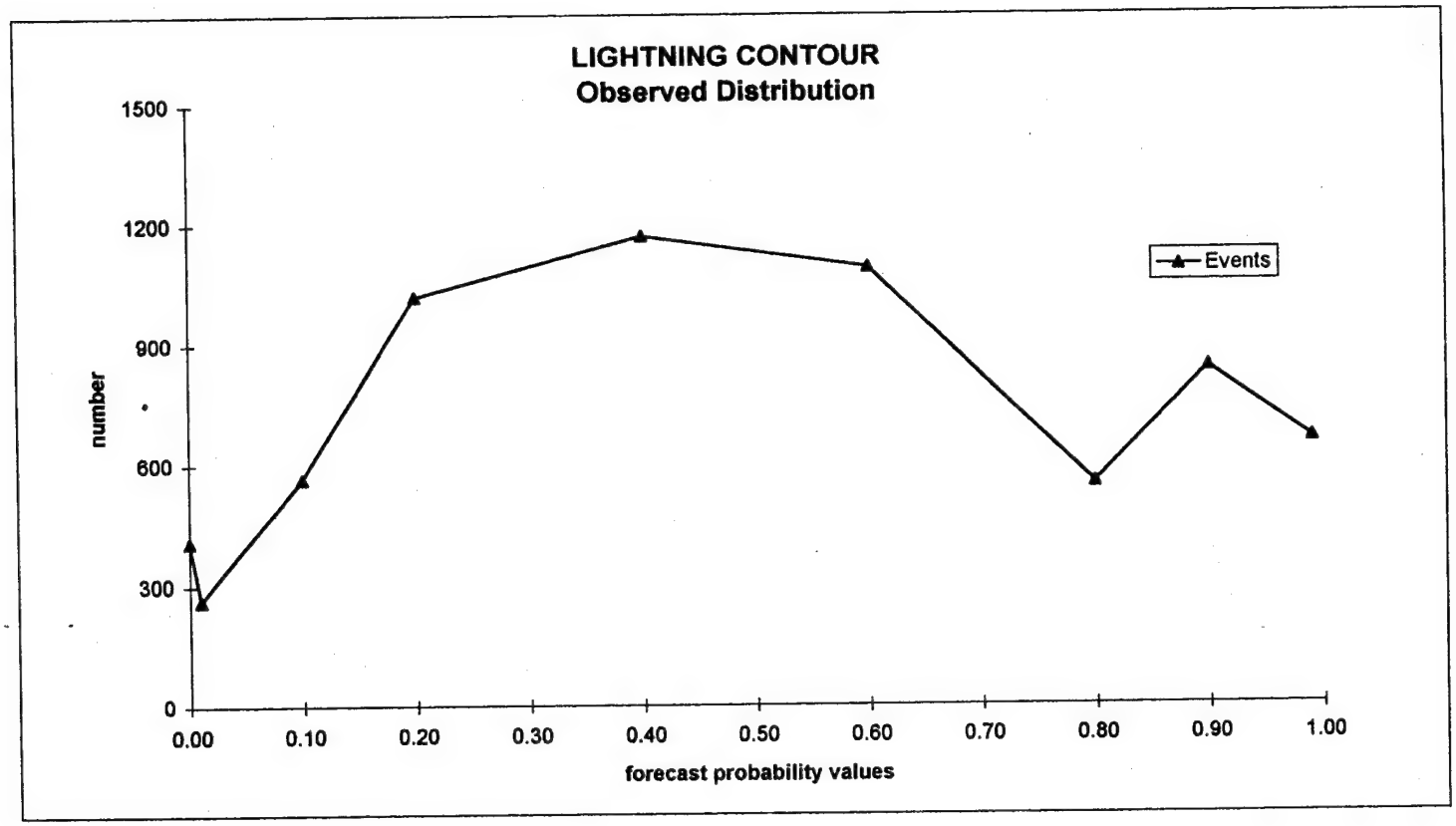


Figure 3b. As in Fig. 3a, except for events only with the scale of the y-axis changed.

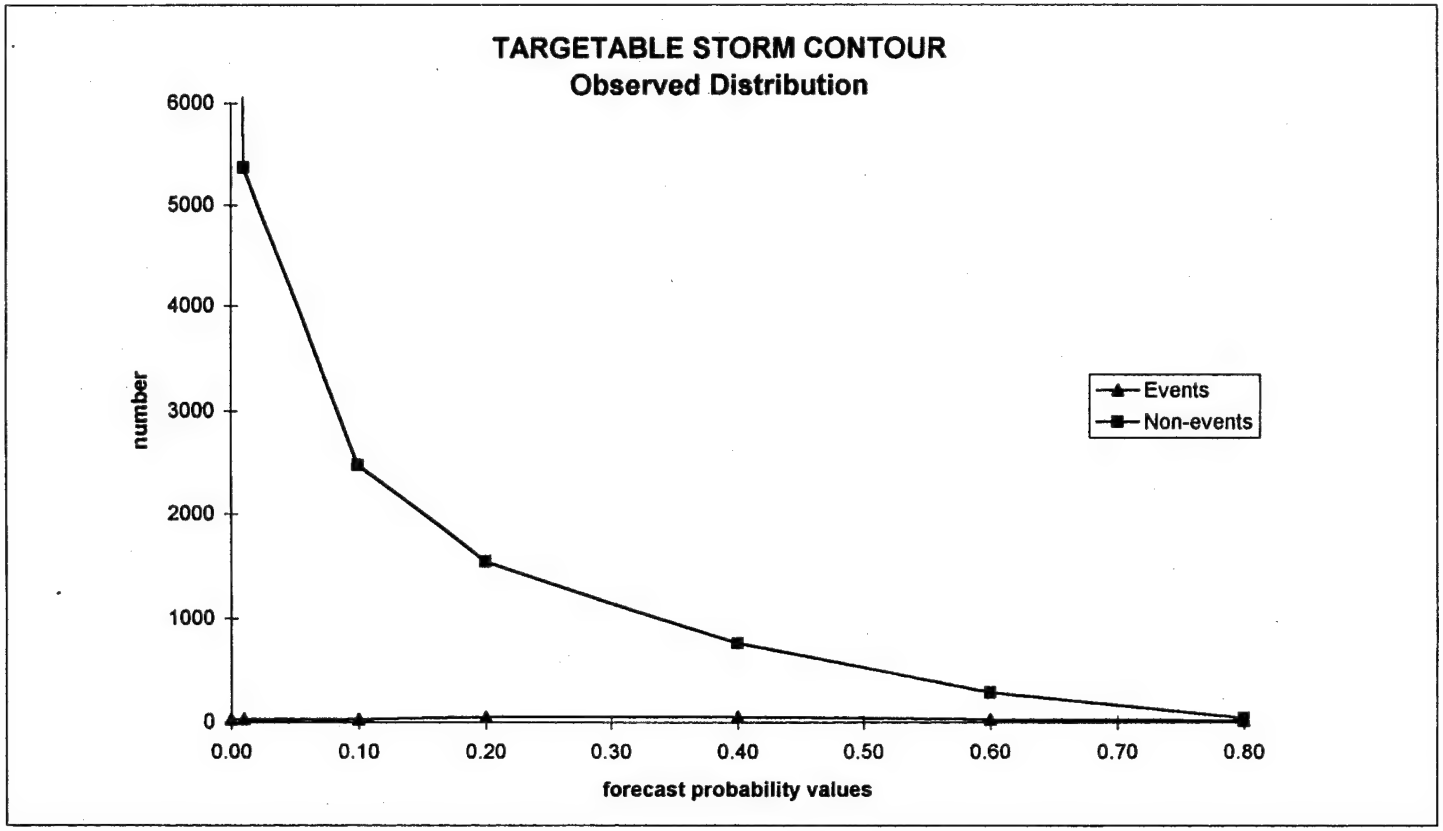


Figure 3c. As in Fig. 3a, except for targetable storm contour forecasts.

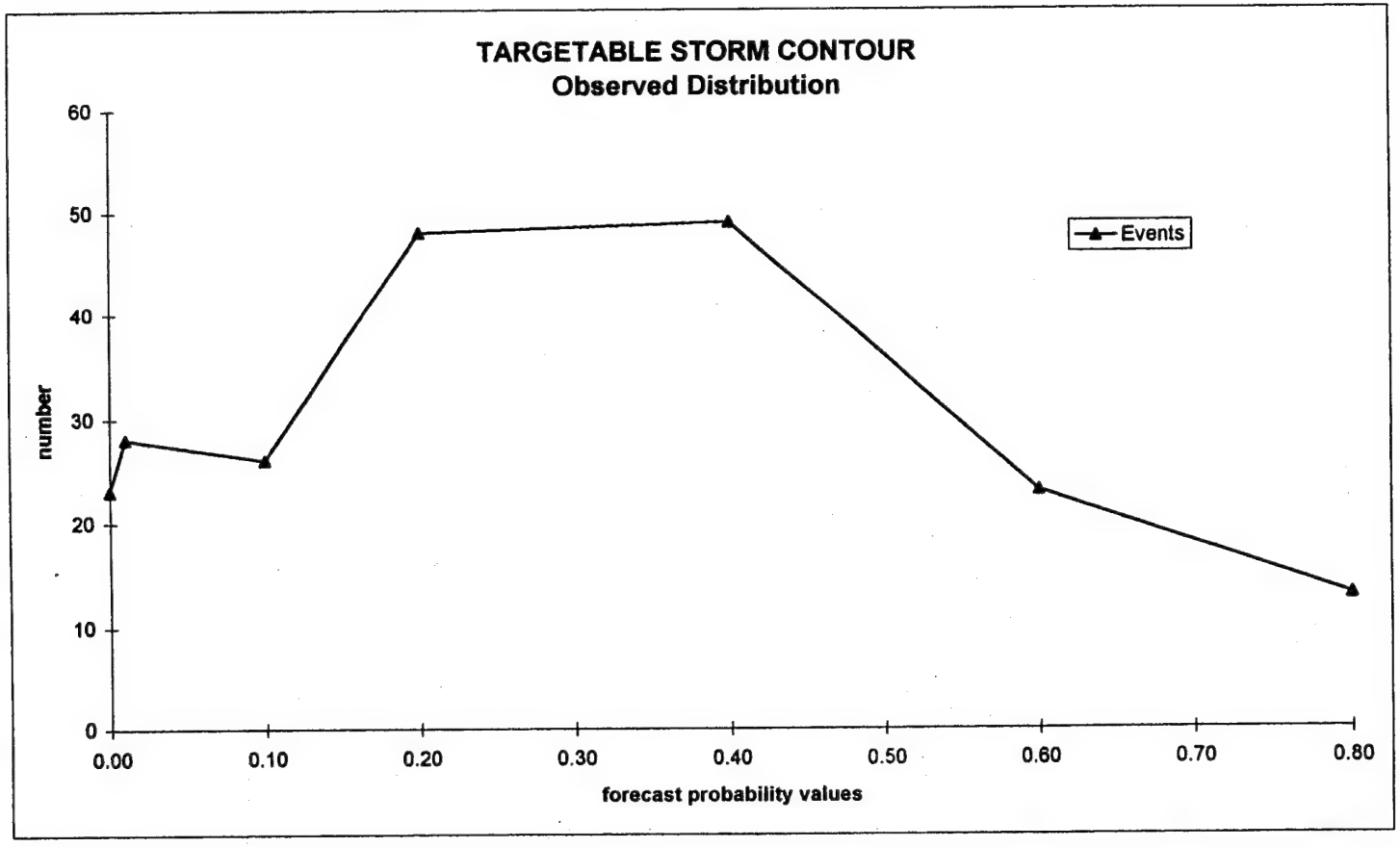


Figure 3d. As in Fig. 3b, except for targetable storm contour forecasts.

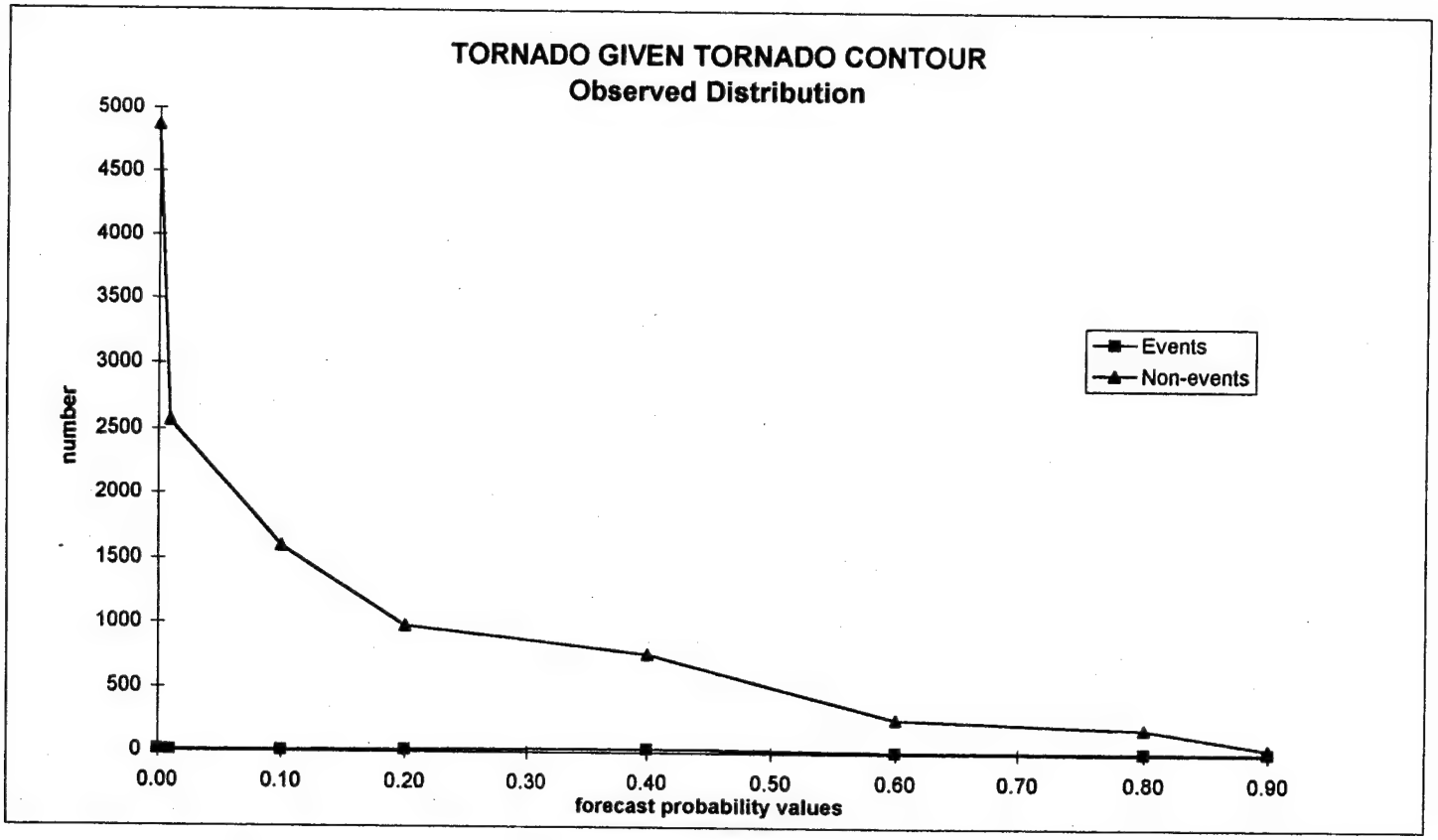


Figure 3e. As in Fig. 3a, except for tornado given tornado contour forecasts.

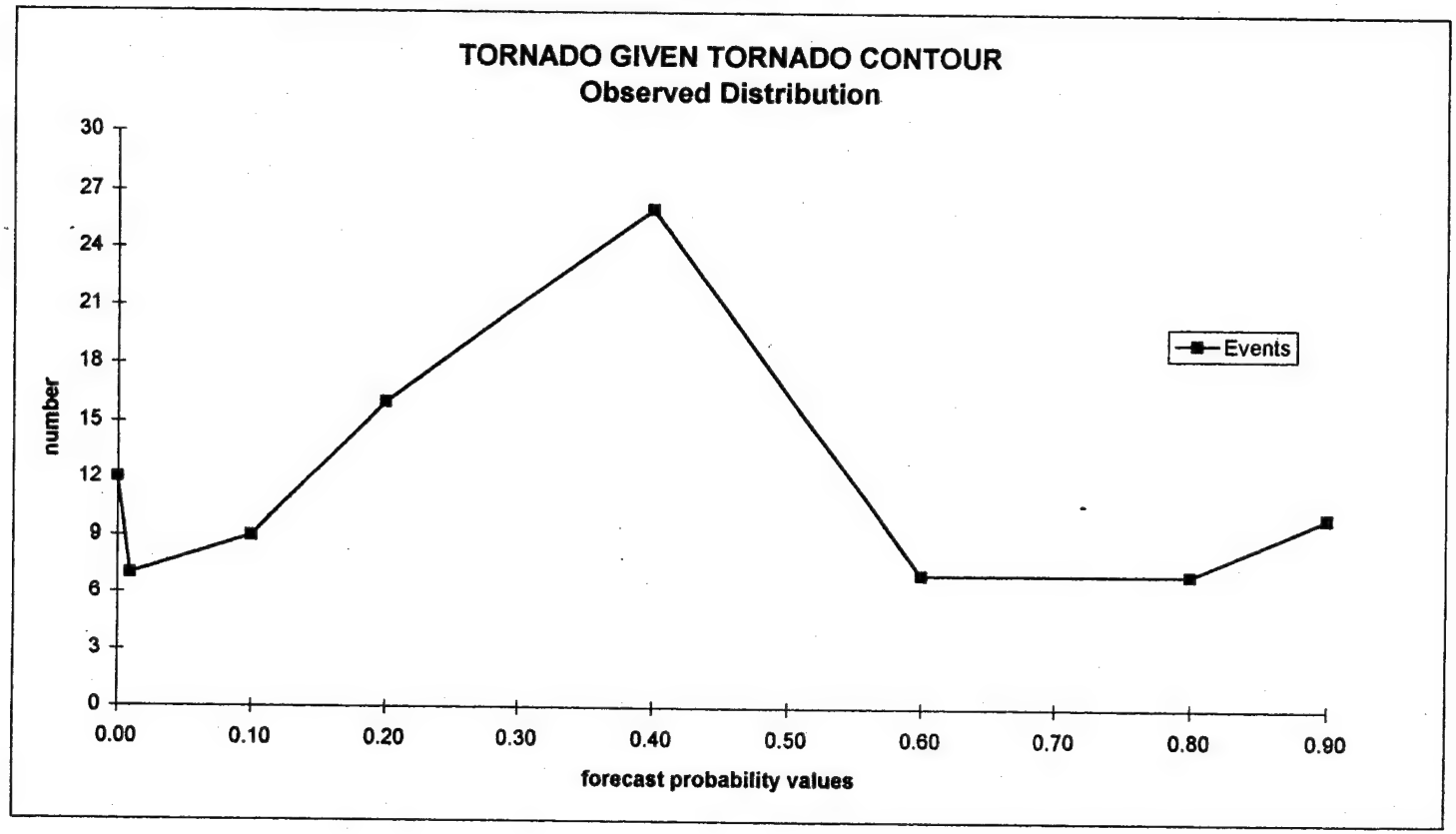


Figure 3f. As in Fig. 3b, except for tornado given tornado contour forecasts.

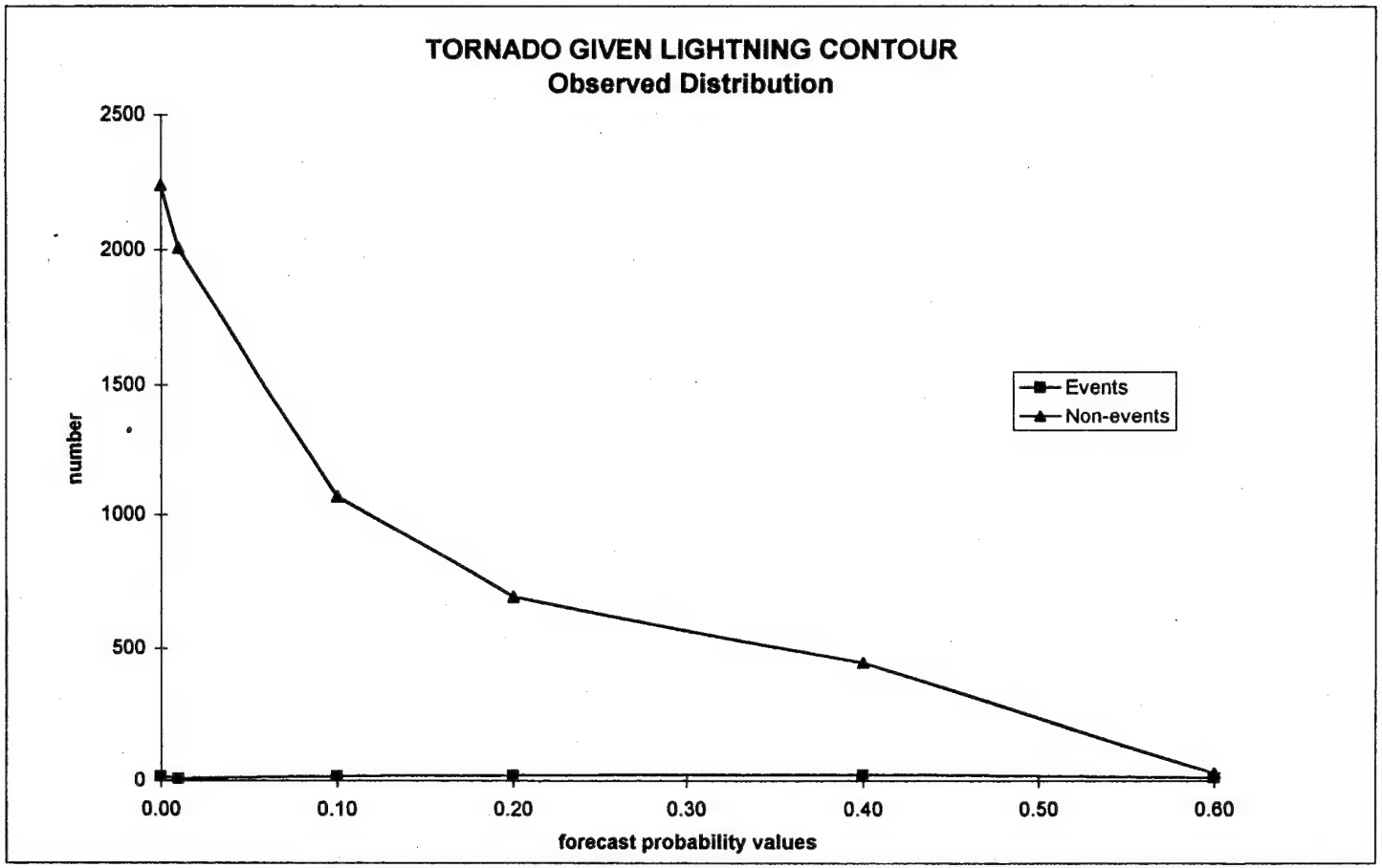


Figure 3g. As in Fig. 3a, except for tornado given lightning contour forecasts.

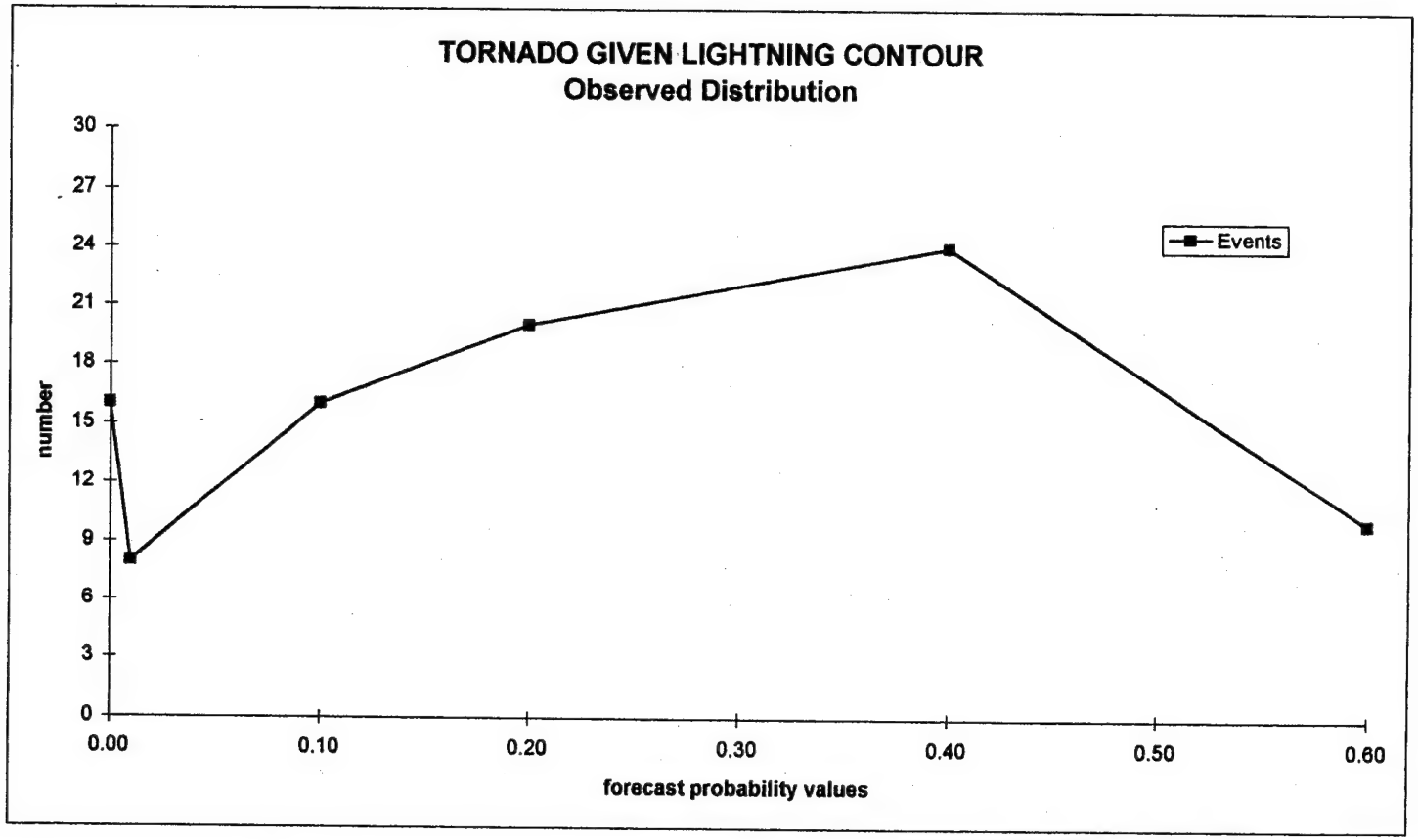


Figure 3h. As in Fig. 3b, except for tornado given lightning contour forecasts.

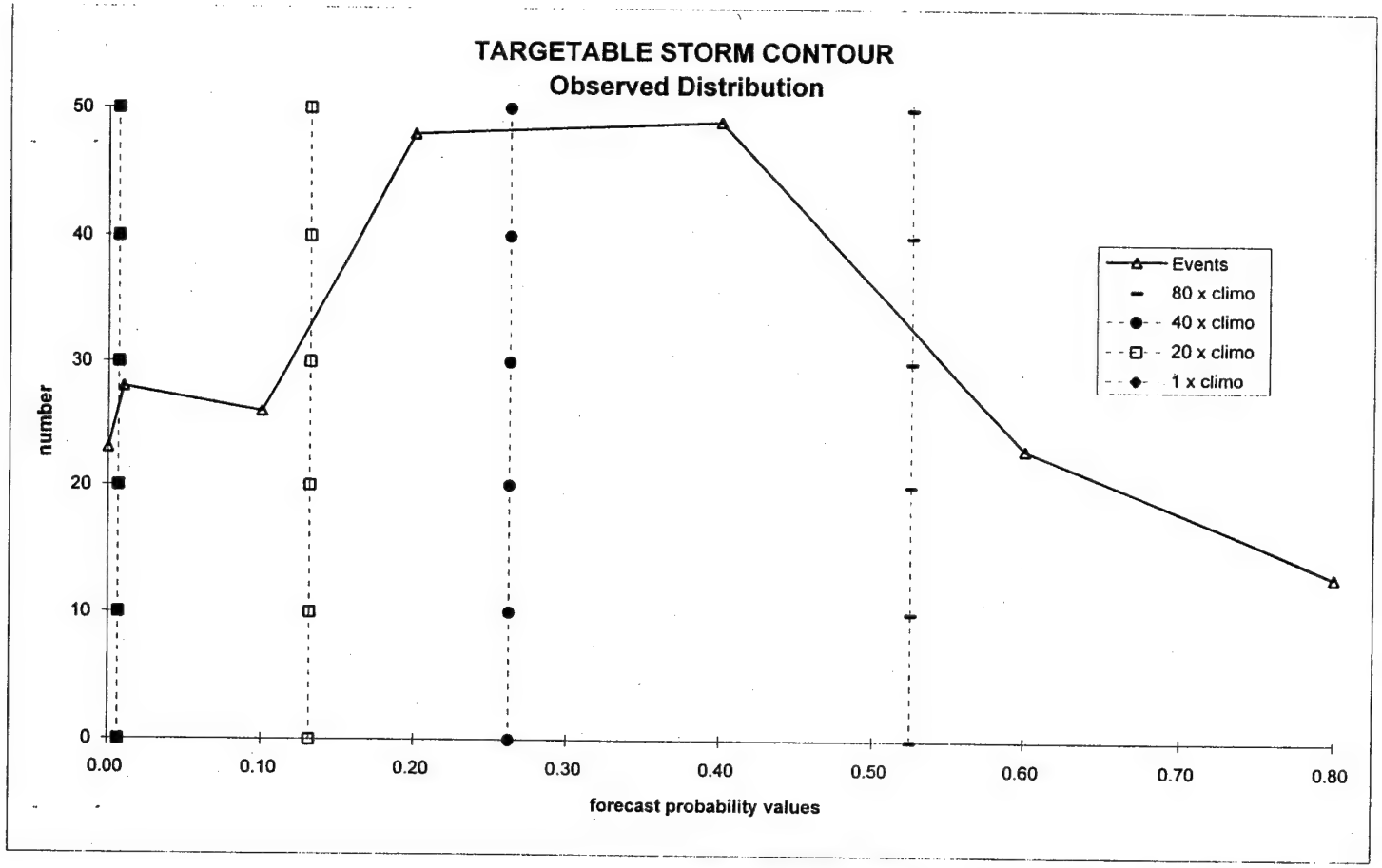


Figure 3i. Observed distribution of events for targetable storm contour forecasts, designating which forecasts will be lumped together by multiples of the climatology.

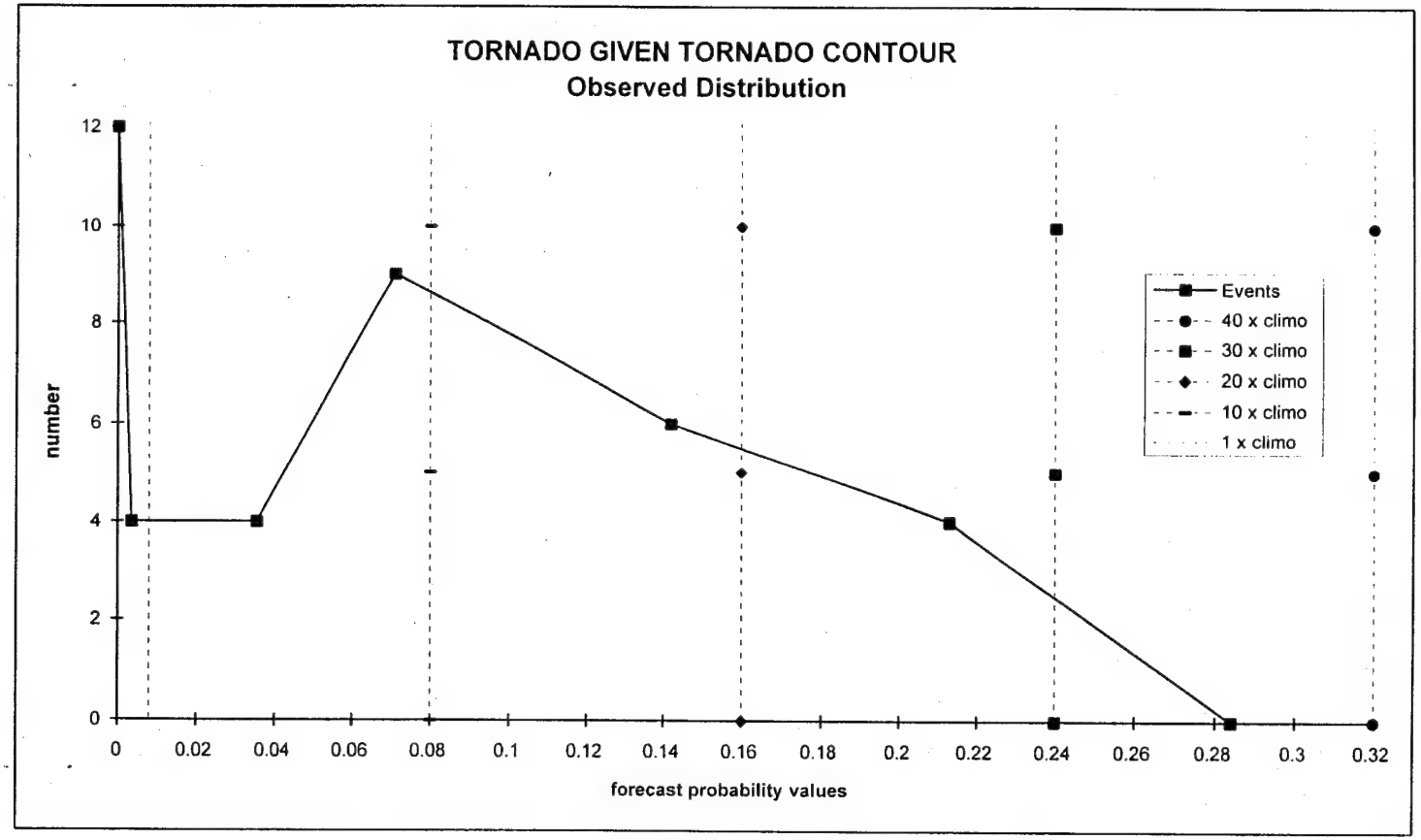


Figure 3j. As in Fig. 3i, except for tornado given tornado contour forecasts.

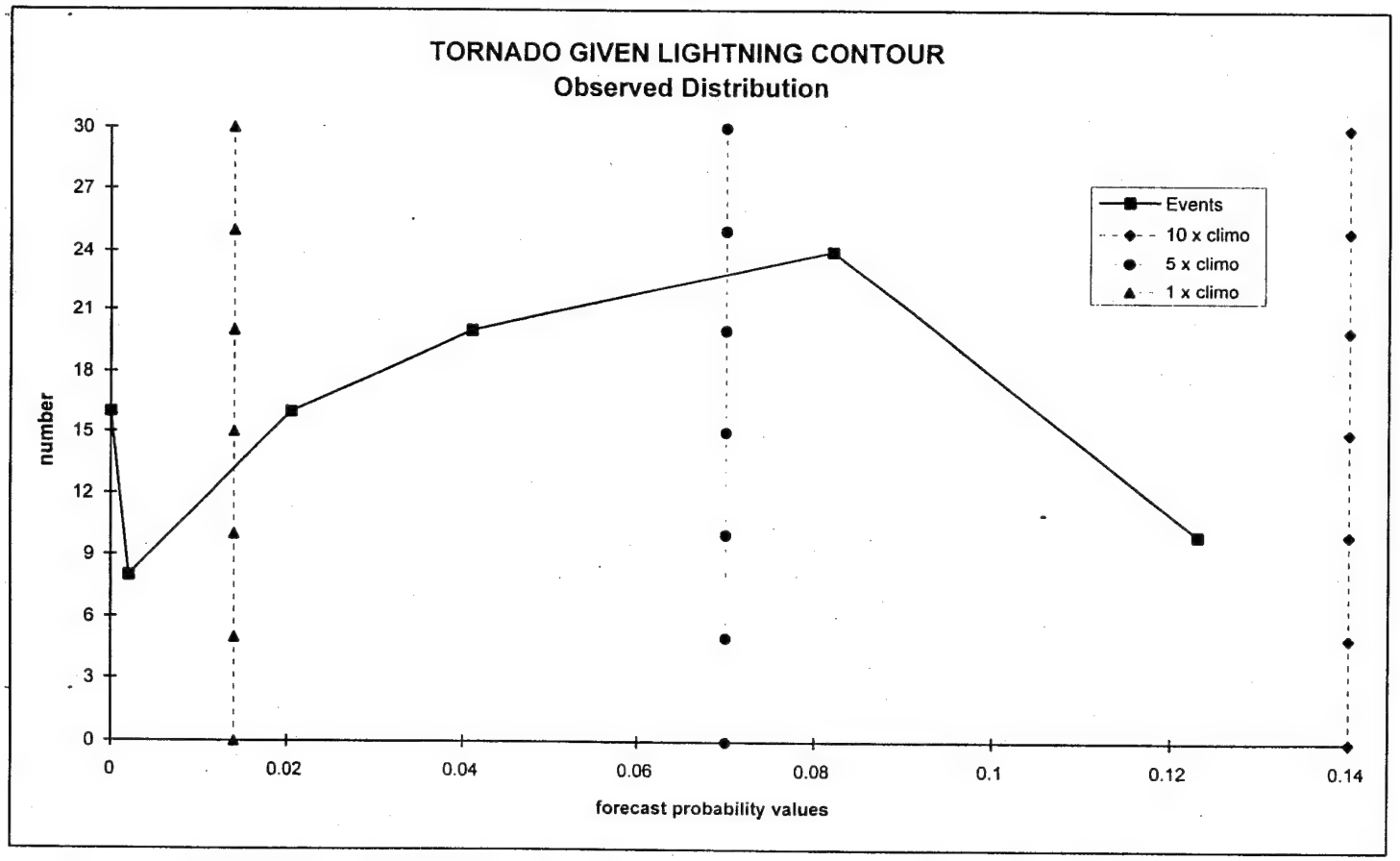


Figure 3k. As in Fig. 3i, except for tornado given lightning contour forecasts.

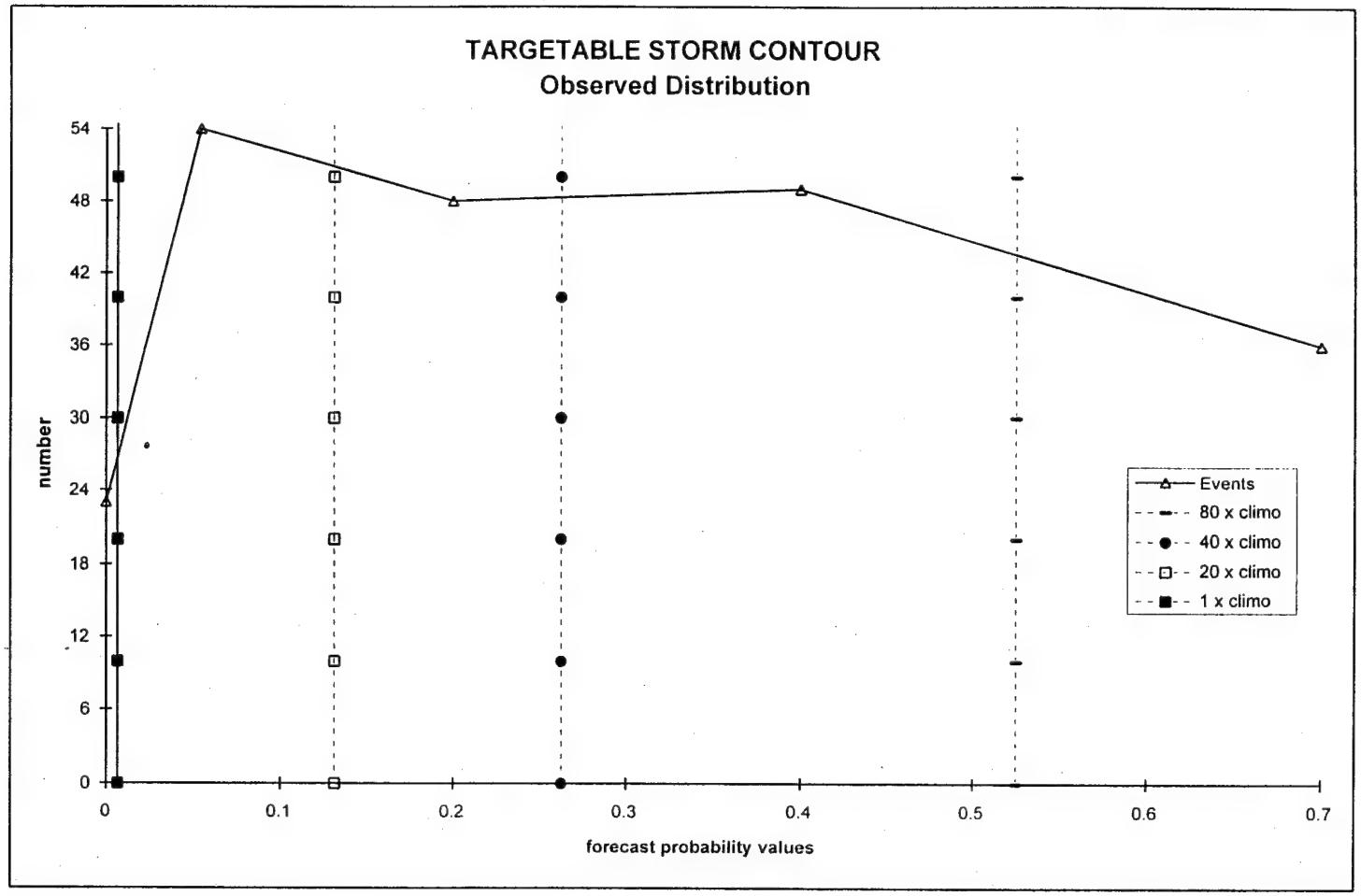


Figure 3l. Observed distribution of events for targetable storm contour forecasts, after they've been re-grouped.

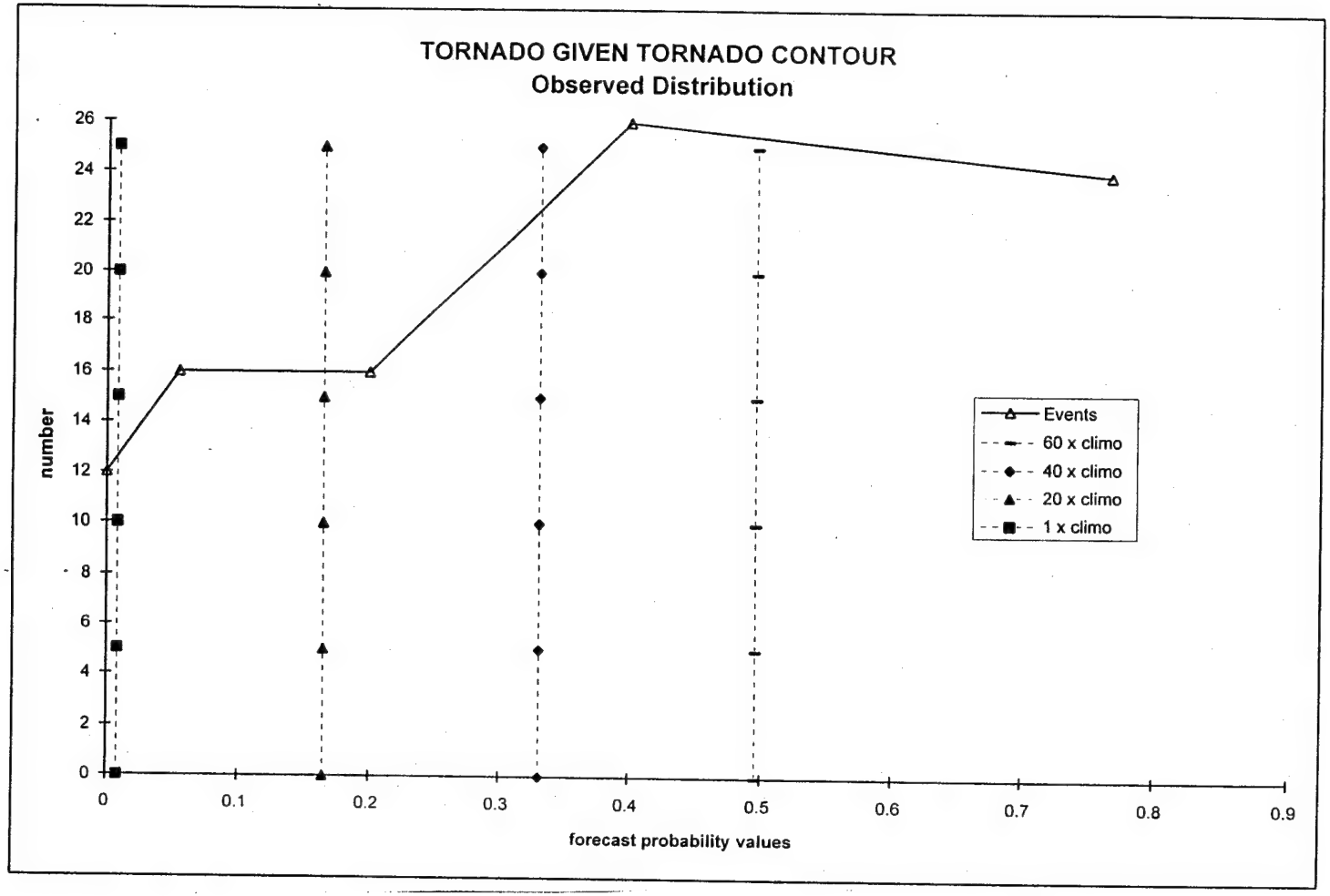


Figure 3m. As in Fig. 3l, except for tornado given tornado contour forecasts.

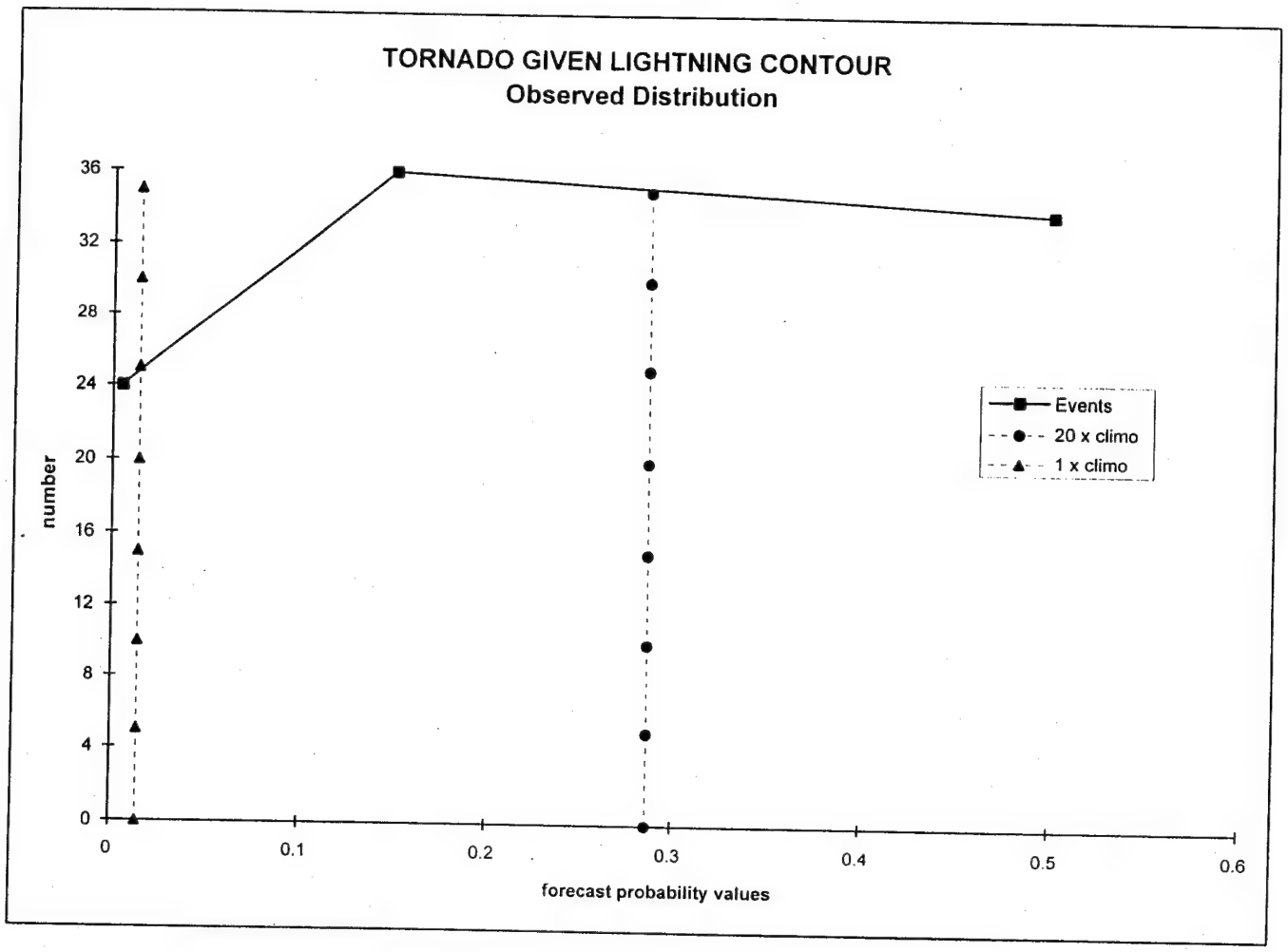


Figure 3n. As in Fig. 3l, except for tornado given lightning contour forecasts.

It is the spatial rarity of these events that makes them a challenge to forecast and verify correctly. One of the ways to deal with this rarity is to decrease the resolution of our forecasts by grouping some of the probability categories together. This re-grouping was done only for the targetable storm and conditional tornado probability forecasts. The re-grouping was based on multiples of the climatological frequency of each phenomenon (Figs. 3i-k). For example, in Fig. 3i the 1% and 10% forecasts (1-20 x climo) were grouped as well as the 60% and 80% forecasts (greater than 80 x climo) with the other probability categories remaining the same. The new re-binned categories are shown in Figs. 3l-n (Tables 2m-p). This gives more hits in each category and smooths the distributions.

4.3 Measures-oriented approach

4.3.1 Traditional statistical formulas

A measures-oriented approach typically uses a limited number of measures of forecast accuracy and skill. Some of the measures that have been used in the past include: probability of detection (POD), probability of false detection (POFD), hit rate (H), false alarm rate (FAR), Heidke skill score (HSS), Kuipers skill score (KSS), true skill statistic (TSS), skill score (SS), critical success index (CSI; also called the threat score (TS)), correlation coefficient (r), bias (b), mean absolute error (MAE), mean square error (MSE), and the Brier score (BS). Those measures used in this verification study are defined below.

The proportion of forecasting occasions that had hits/events observed is the "hit frequency" (ν_h). Note that this is **not** the same as the hit rate which is the ratio of correct forecasts/total forecasts. "Hit frequency" is given by the following for area forecasts:

$$\nu_H = \frac{H}{n}, \tag{4}$$

where H is the total number of hits and n is the total number of forecasts. For contour forecasts, the total number of hits is the number of grid boxes hit and the total number of forecasts is the total number of grid boxes possible. The proportion of forecasting occasions on which non-events were observed is the "none frequency" (ν_N), which is given by the following for area forecasts:

$$\nu_N = \frac{N}{n}, \tag{5}$$

where N is the total number of non-events and n is the total number of forecasts. For contour forecasts, the total number of non-events is the number of grid boxes not hit and the total number of forecasts is the total number of grid boxes possible. The "average forecast" is the sum of the forecasts in each probability category divided by the total number of forecasts. This is given by the following:

$$\overline{F} = \frac{\sum_i D_i F_i}{n}, \tag{6}$$

where F_i is the forecast probability for the category i, D_i is the number of times forecast category i was used, and n is the total number of forecasts. The variance of the forecasts about the mean forecast is termed the "forecast variance". This is given by the following:

$$VAR(F) = \frac{\sum_i D_i (F_i - \overline{F})^2}{n}, \tag{7}$$

where D_i, F_i, and n were previously defined. The variance of the observed events about their mean is termed the "event variance":

$$VAR(X) = \frac{H(1-\overline{X})^2 + N(0-\overline{X})^2}{n}, \tag{8}$$

where H, N, and n were previously defined and the base rate ($\overline{X}$) is given by:

$$\overline{X} = \nu_H = \frac{H}{n}. \tag{9}$$

The "bias" is defined as the difference between the average forecast and average observation, so a value near zero indicates unbiased forecasting:

$$b = \overline{F} - \overline{X}. \tag{10}$$

The most common method for evaluating the accuracy of probability forecasts is the "Brier score" (BS). This is essentially the mean-squared error (MSE) of the probability forecasts. It can take on values between 0 and 1, with 0 indicating a perfect forecast. The general form of the Brier score is:

$$BS = \frac{\sum_i (F_i - X_i)^2}{n}, \tag{11}$$

where X_i is the observation and F_i is the forecast for probability category i. Extending this basic definition to the two possible observations, X = 1 and X = 0, results in the following formula:

$$BS = \frac{\sum_i \left(H_i (F_i - 1)^2 + N_i (F_i - 0)^2 \right)}{n}, \tag{12}$$

where H_i is the number of hits in category i and N_i is the number of non-events in category i. The "skill score" (SS) is a measure of the skill of our forecasts measured

by how much of an improvement is noted over some baseline forecast (e.g. sample climatology). The skill score can take on values between 0 and 1, with 1 being a 100% improvement over climatology. In its most generic form, the skill score is given by:

$$SS = \frac{X - X_r}{X_p - X_r} \tag{13}$$

where X is some measure of accuracy which, for this study, is the Brier score, X_p is the value of X for a perfect forecast which in this case is zero, and X_r is the value of X for the reference forecast which in this case is climatology. The skill score then, in this case, is given by the following:

$$SS = 1 - \frac{BS}{BS_c}, \tag{14}$$

$$BS_c = \frac{H}{n}(\overline{X} - 1)^2 + \frac{N}{n}(\overline{X} - 0)^2 \quad , \tag{15}$$

where BS_c is the Brier score of a climatological forecast. The basic definitions for these measures can be found in Wilks (1995).

4.3.2 Results using traditional statistics (Day-1 and Day-2 forecasts)

Some of the traditional measures of accuracy and skill are presented in Table 3. The hit frequencies increased slightly for Day-2 forecasts because of the larger time window and because there are now only 71 forecast days for Day-2 forecasts vs. 76 forecast days for Day-1 forecasts. Considering the size of the VORTEX area, the high hit frequencies are not all that surprising. In an area of this size, it is quite likely that CG lightning and severe weather will be observed on any given day of the spring.

The Brier score shows lightning to be the most accurate with a score of .12. This Brier score is in the same range as typical NWS PoP forecasts (.07-.13). Typical skill scores for these forecasts are from 27-56%, significantly higher than the 13% obtained here (Dagostaro et al. 1995). For severe forecasts, the Brier score is .18 with a skill score of 20%. This is both less accurate and less skillful than in past field experiments. During DOPLIGHT '87, the noon outlook gave the probability of seeing severe weather within 230 km of the Norman, Oklahoma radar. The Brier score for the DOPLIGHT noon outlook was .12, with about a 40% improvement over climatology (Doswell and Flueck 1989). This same Brier score, .12, was also obtained during MAP '88/'89 for both the noon outlook (SS = 43%) and the advance outlook (SS = 23%) (Jincai et al. 1992). These experiments also had less underforecasting: Doplight '87 (bias = -11%); MAP '88/'89 (Day-1 bias = -3%, Day-2 bias = 0); VORTEX '94 (bias = -20%). For tornado forecasts, the Brier score of .21 is comparable to the scores obtained during a probabilistic forecasting experiment at the

National Severe Storms Forecast Center where they forecast the probability of seeing a tornado within an outlook area (BS = .20) and within a watch area (BS = .24). The skill scores for this experiment were 19% and -2% for outlook and watch areas, respectively (Murphy and Winkler 1982). Thus, forecasting the probability of tornado occurrence within a SELS outlook area (SS = 19%) was more skillful than forecasting the probability of tornado occurrence within the VORTEX '94 area (SS = 9%). Note, however, that the outlook area is moved around to cover the highest risk areas whereas the VORTEX area is fixed in its location.

Though it appears that these past experiments were more accurate, skillful, and less biased, the results must be used with caution. The forecasts made during DOPLIGHT '87, MAP '88/'89, and VORTEX '94 are hard to compare directly because of the issue times of the forecasts (1200L (DOPLIGHT '87 and MAP '88/'89) vs. 0900L (VORTEX '94)), the different sizes of the forecast areas, and because the severe reports in these different experiments were verified by radar (DOPLIGHT '87) or severe warnings (MAP '88/'89) vs. actual reports (VORTEX '94). It is also hard to compare these forecasts directly to the precipitation probability forecasts issued by the National Weather Service or the SELS tornado probability forecasts because of the different sized forecast areas, different weather phenomena forecast, and different methods of verification (Doswell and Flueck 1989, Jincai et al. 1992, Murphy and Winkler 1982).

4.3.3 Results using traditional statistics (Day-1 graphical probability forecasts)

The same measures of accuracy and skill presented above also have been computed for the contour forecasts (Table 3). The hit frequencies are very small, especially for targetable storms and tornadoes, because they represent the fraction of grid boxes hit out of all grid boxes possible for a given day. The hit frequency is higher for TGL than for TGT because the same number of tornadoes is occurring over fewer possible grid boxes for TGL.

It might be expected that the forecasts would decrease in accuracy as the events become rarer but, looking at the numbers, it appears that the opposite is true. The Brier scores seem to indicate that the targetable storm and conditional tornado probability forecasts did rather well, but these numbers are misleading. These numbers are heavily weighted toward zero by the large number of correctly forecast non-events. The Brier score then, is a poor measure of accuracy for rare events (Jincai et al. 1992). Since the skill score uses the Brier score in its computation, the skill scores are also unreliable for the targetable storm and conditional tornado forecasts. In contrast to the area forecasts, the contour forecasts are characterized by overforecasting of 2-9%, depending upon the phenomenon. The conditional tornado probability forecasts were overforecast the most, the result of the small spatial coverage of the events compared to the much larger spatial coverage of the contour areas.

Table 3. Measures of accuracy and skill for all forecasts.

	Day-1					Day-2					Contour		
Measure	L	S	T	TS		L	S	T	TS		L	TS	TGT	TGL
hit frequency	.84	.65	.36	.42		.85	.65	.37	.44		.21	.007	.008	.014
avg forecast	.68	.45	.19	.24		.64	.42	.18	.22		.23	.04	.10	.07
base rate	.84	.65	.36	.42		.85	.65	.37	.44		.21	.008	.008	.014
bias	-.16	-.20	-.16	-.18		-.21	-.23	-.18	-.22		.02	.03	.09	.07
Brier score	.12	.18	.21	.22		.16	.22	.23	.23		.14	.01	.02	.01
Brier score (climo)	.13	.23	.23	.24		.13	.23	.23	.25		.16	.007	.008	.014
Skill Score	.13	.20	.09	.12		-.21	.05	-.01	.07		.15	-	-	-

These traditional measures (Brier score, skill score, and bias) don't give enough information to describe fully the quality of the forecasts. The Brier scores, for example, indicate which forecasts are the most accurate, but don't indicate what parts of the forecast range are more or less accurate than others. The skill scores indicate whether there has been some improvement over climatology, but they don't identify what parts of the forecast range are more or less skillful than others. The bias indicates the degree of under- or overforecasting, but it still doesn't indicate where the under- or overforecasting is the worst. To provide more insight into forecast quality, then, the distributions-oriented approach must be used.

4.4 Distributions-oriented marginal distributions ($p(f)$ and $p(x)$)

4.4.1 VORTEX area forecasts

The marginal distribution of the observations (Tables 4a,b) identifies the base rate (sample climatology) of events and non-events. These are equivalent to the "hit" frequency and "none" frequency presented in the previous section. The marginal distribution of forecasts, $p(f)$, tells us how often different forecast values are used. The marginal distribution of the forecasts is also presented graphically in Figs. 4a-d and 5a-d for each of the area forecasts. Ideally, for well refined forecasts, most of the forecasts should be close to zero or one with the middle probabilities used less often. However, this ideal is only obtainable as the state-of-the-art in forecasting these phenomena improves. In this experiment, lower probabilities were used more often

than higher probabilities, as expected, as the events became rarer or, as the forecasts went to longer ranges (e.g., Day-1 to Day-2 forecasts). This can be easily seen by just looking at the Day-1 forecasts (Figs. 4a-d). The most common forecast was, for: lightning, 100%; severe weather, 50%; tornadoes, 1%; and targetable storms, 0%. The rarer the event being forecast, then, the less refined the forecasts will be.

4.4.2 VORTEX contour forecasts

The marginal distribution of the forecasts, $p(f)$, and observations, $p(x)$, for the contour forecasts are shown in Table 4c-d (Figs. 6a-d). Note that successively lower probabilities were used as the events became rarer. The zero and one percent probabilities were used more often than any other value for all of the contour forecasts. The small usage of higher probabilities is understandable and clearly is associated with the rarity of the events.

4.5 Distributions-oriented conditional distributions $p(x \mid f)$, $p(f \mid x)$

The conditional probability distributions are presented in Tables 5a-d and 6a-d. It was previously noted that the $p(f \mid x)$ is related to the discrimination while the $p(x \mid f)$ is related to both the reliability and the resolution. These quality indicators are presented both numerically and graphically in the discussion that follows.

Table 4a. Percentage of forecasts in each category for Day-1 area forecasts.

	Lightning			Severe			Tornadoes			Targetable storms		
Prob(%)	H	N	P(F)	H	N	P(F)	H	N	P(F)	H	N	P(F)
0	0.0	2.6	2.6	0.0	7.9	7.9	0.0	13.2	13.2	1.3	14.5	15.8
2	0.0	2.6	2.6	0.0	2.6	2.6	3.9	15.8	19.7	0.0	11.8	11.8
5	0.0	3.9	3.9	1.3	6.6	7.9	2.6	7.9	10.5	3.9	9.2	13.2
10	2.6	2.6	5.3	5.3	5.3	10.5	5.3	7.9	13.2	6.6	5.3	11.8
20	3.9	0.0	3.9	1.3	1.3	2.6	2.6	9.2	11.8	3.9	6.6	10.5
30	1.3	0.0	1.3	5.3	3.9	9.2	9.2	5.3	14.5	5.3	5.3	10.5
40	3.9	1.3	5.3	7.9	1.3	9.2	1.3	0.0	1.3	3.9	1.3	5.3
50	9.2	1.3	10.5	9.2	3.9	13.2	1.3	3.9	5.3	3.9	0.0	3.9
60	2.6	0.0	2.6	5.3	2.6	7.9	3.9	1.3	5.3	2.6	2.6	5.3
70	5.3	1.3	6.6	6.6	0.0	6.6	3.9	0.0	3.9	2.6	0.0	2.6
80	7.9	0.0	7.9	5.3	0.0	5.3	0.0	0.0	0.0	6.6	1.3	7.9
90	2.6	0.0	2.6	6.6	0.0	6.6	0.0	0.0	0.0	0.0	0.0	0.0
95	9.2	0.0	9.2	3.9	0.0	3.9	1.3	0.0	1.3	1.3	0.0	1.3
98	2.6	0.0	2.6	1.3	0.0	1.3	0.0	0.0	0.0	0.0	0.0	0.0
100	32.9	0.0	32.9	5.3	0.0	5.3	0.0	0.0	0.0	0.0	0.0	0.0
P(X)	84	16	100	65	35	100	35	65	100	42	58	100

Table 4b. Percentage of forecasts in each category for Day-2 area forecasts.

	Lightning				Severe				Tornadoes				Targetable storms		
Prob (%)	H	N	P(F)		H	N	P(F)		H	N	P(F)		H	N	P(F)
0	0.0	1.4	1.4		0.0	4.2	4.2		0.0	9.9	9.9		1.4	9.9	11.3
2	0.0	2.8	2.8		0.0	5.6	5.6		4.2	12.7	16.9		1.4	9.9	11.3
5	1.4	1.4	2.8		2.8	4.2	7.0		2.8	9.9	12.7		5.6	5.6	11.3
10	2.8	2.8	5.6		4.2	4.2	8.5		4.2	9.9	14.1		0.0	11.3	11.3
20	4.2	2.8	7.0		2.8	2.8	5.6		9.9	7.0	16.9		9.9	8.5	18.3
30	7.0	0.0	7.0		7.0	4.2	11.3		4.2	5.6	9.9		8.5	8.5	16.9
40	1.4	1.4	2.8		8.5	4.2	12.7		4.2	7.0	11.3		2.8	2.8	5.6
50	9.9	0.0	9.9		11.3	2.8	14.1		1.4	1.4	2.8		4.2	0.0	4.2
60	2.8	1.4	4.2		2.8	1.4	4.2		2.8	0.0	2.8		1.4	0.0	1.4
70	5.6	0.0	5.6		8.5	1.4	9.9		2.8	0.0	2.8		4.2	0.0	4.2
80	7.0	1.4	8.5		7.0	0.0	7.0		0.0	0.0	0.0		4.2	0.0	4.2
90	15.5	0.0	15.5		8.5	0.0	8.5		0.0	0.0	0.0		0.0	0.0	0.0
95	7.0	0.0	7.0		1.4	0.0	1.4		0.0	0.0	0.0		0.0	0.0	0.0
98	7.0	0.0	7.0		0.0	0.0	0.0		0.0	0.0	0.0		0.0	0.0	0.0
100	12.7	0.0	12.7		0.0	0.0	0.0		0.0	0.0	0.0		0.0	0.0	0.0
P(X)	84	16	100		65	35	100		37	63	100		44	65	100

Table 4c. Percentage of forecasts in each category for lightning and targetable storm contour forecasts.

	L					TS		
Prob (%)	H	N	P(F)		Prob (%)	H	N	P(F)
0	1.3	28.3	29.5		0	0.1	66.7	66.7
1	.8	14.8	15.6		5.5	0.2	24.5	24.7
10	1.8	10.6	12.3		20	0.2	4.8	5.0
20	3.2	9.8	13.0		40	0.2	2.4	2.5
40	3.7	6.9	10.5		70	0.1	1.0	1.1
60	3.4	5.3	8.7					
80	1.7	2.0	3.7					
90	2.6	1.4	4.0					
99	2.1	0.5	2.6					
P(X)	20.5	79.5	100		P(X)	0.7	99.3	100

Table 4d. Percentage of forecasts in each category for conditional tornado probability forecasts.

	TGT					TGL		
Prob (%)	H	N	P(F)		Prob (%)	H	N	P(F)
0	0.1	42.8	42.9		.5	0.4	64.6	65.0
5.5	0.1	36.6	36.7		15	0.5	26.8	27.3
20	0.1	8.6	8.7		50	0.5	7.2	7.7
40	0.2	6.7	6.9					
77	0.2	4.5	4.7					
P(X)	.8	99.2	100		P(X)	1.4	98.6	100

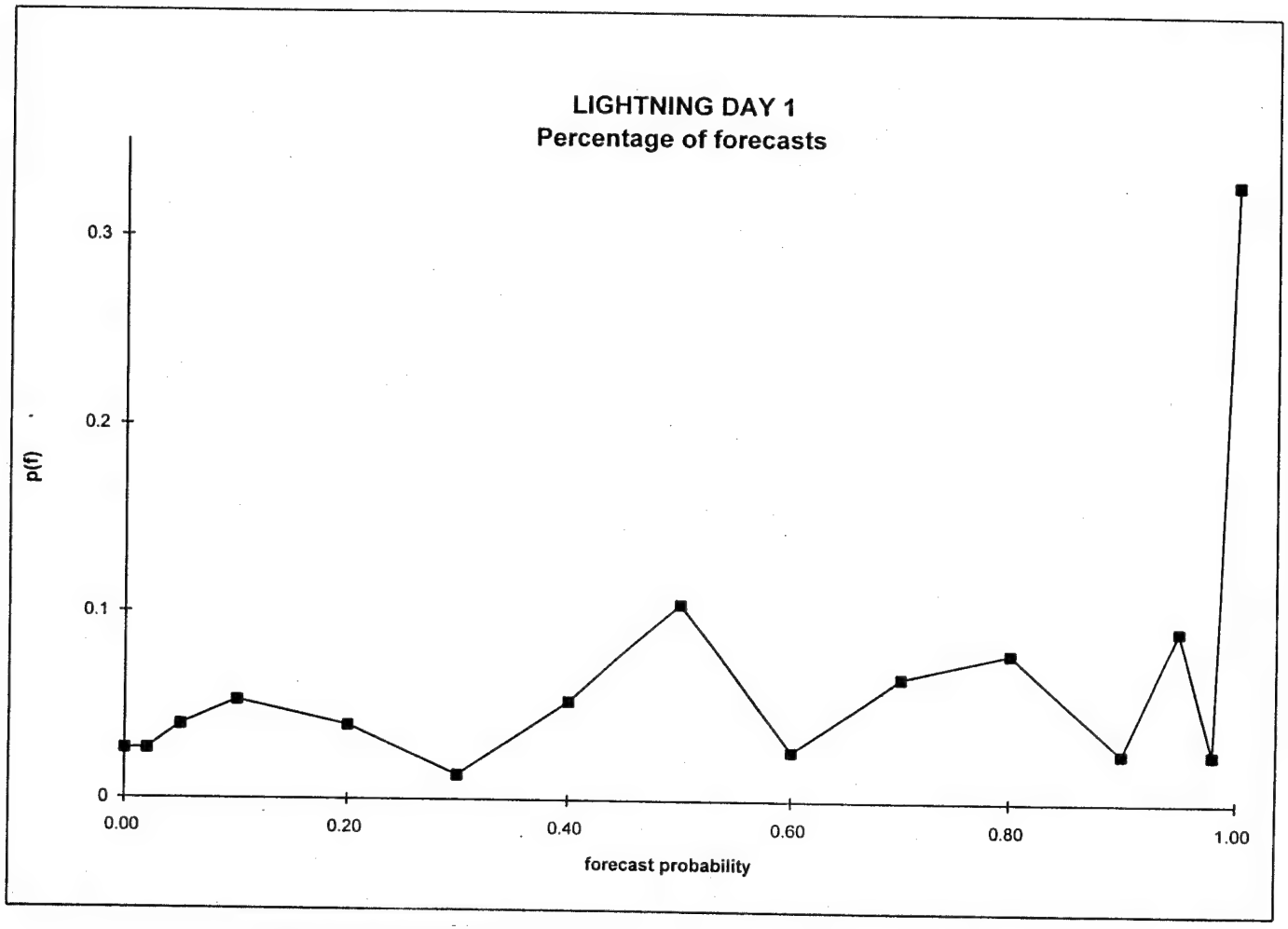


Figure 4a. Percentage of forecasts in each category for lightning Day-1 forecasts.

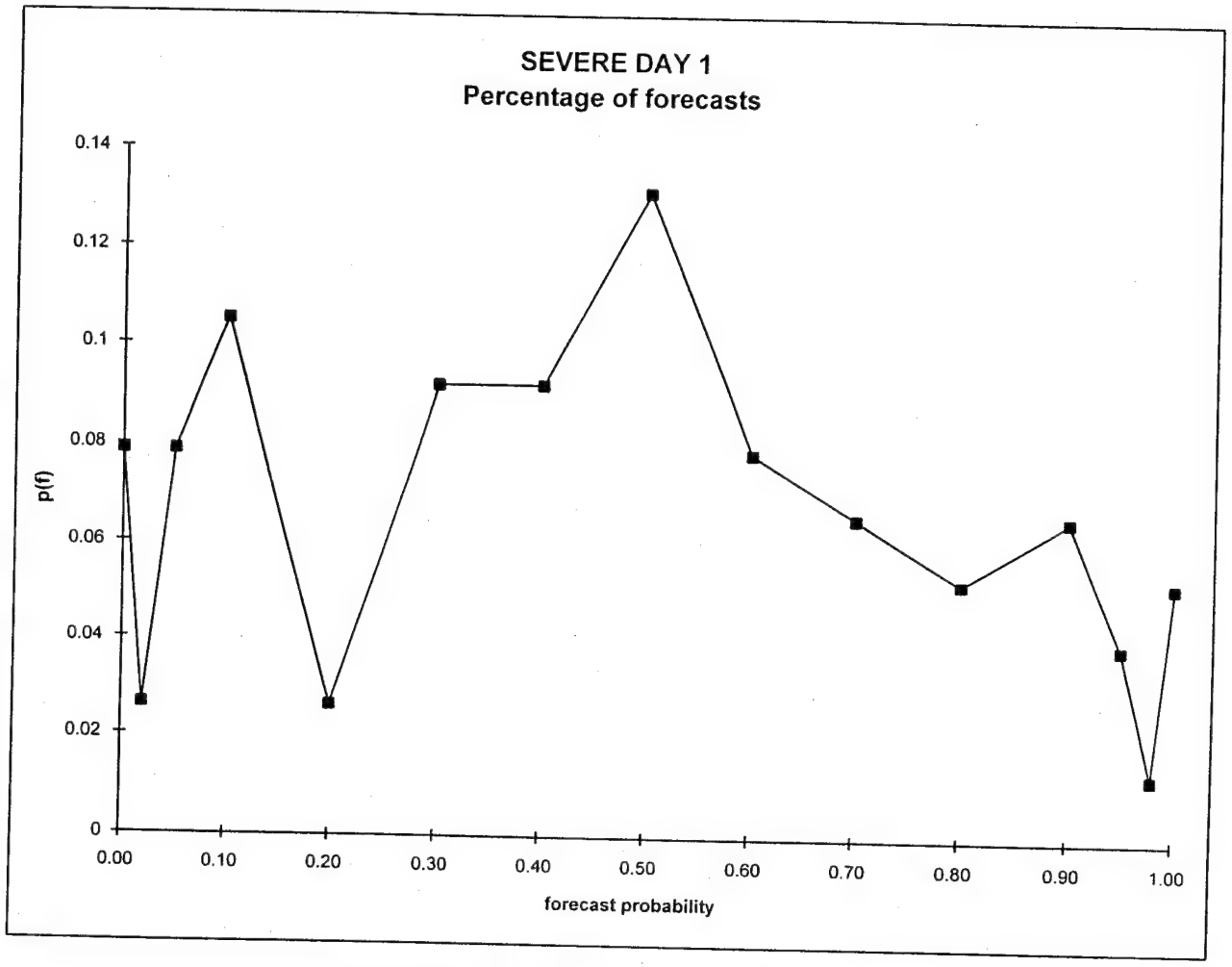


Figure 4b. As in Fig. 4a, except for severe Day-1 forecasts.

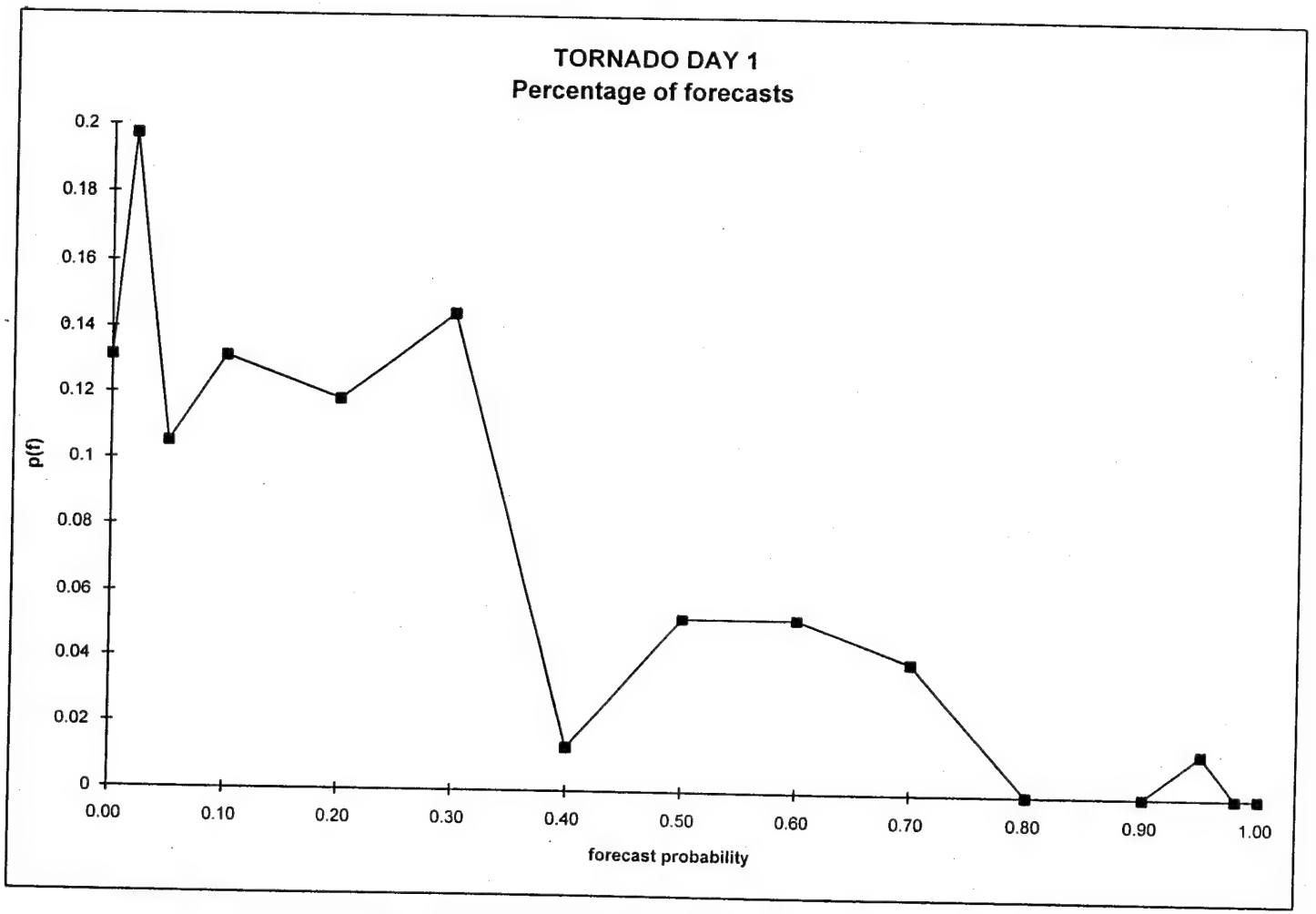


Figure 4c. As in Fig. 4a, except for tornado Day-1 forecasts.

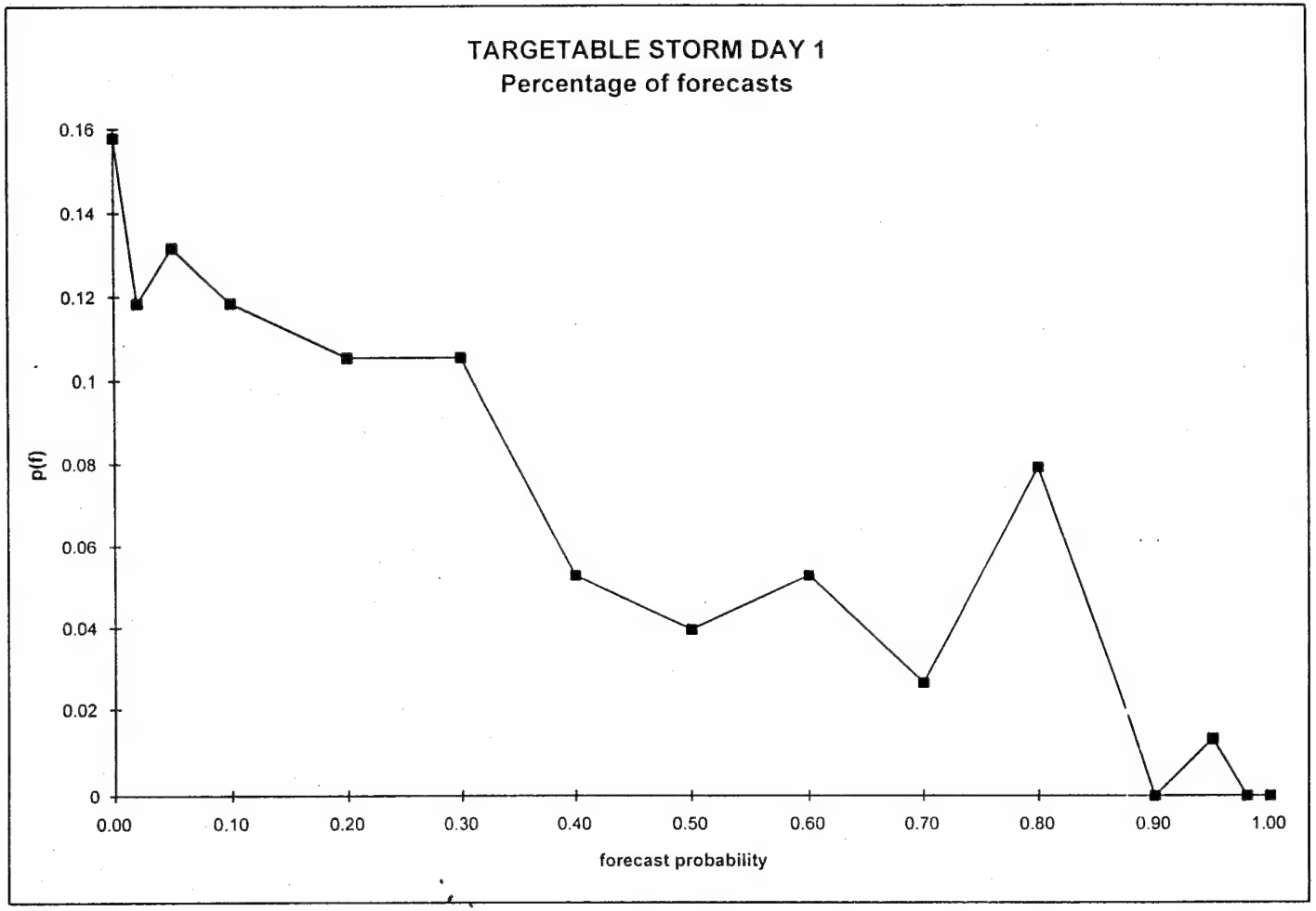


Figure 4d. As in Fig. 4a, except for targetable storm Day-1 forecasts.

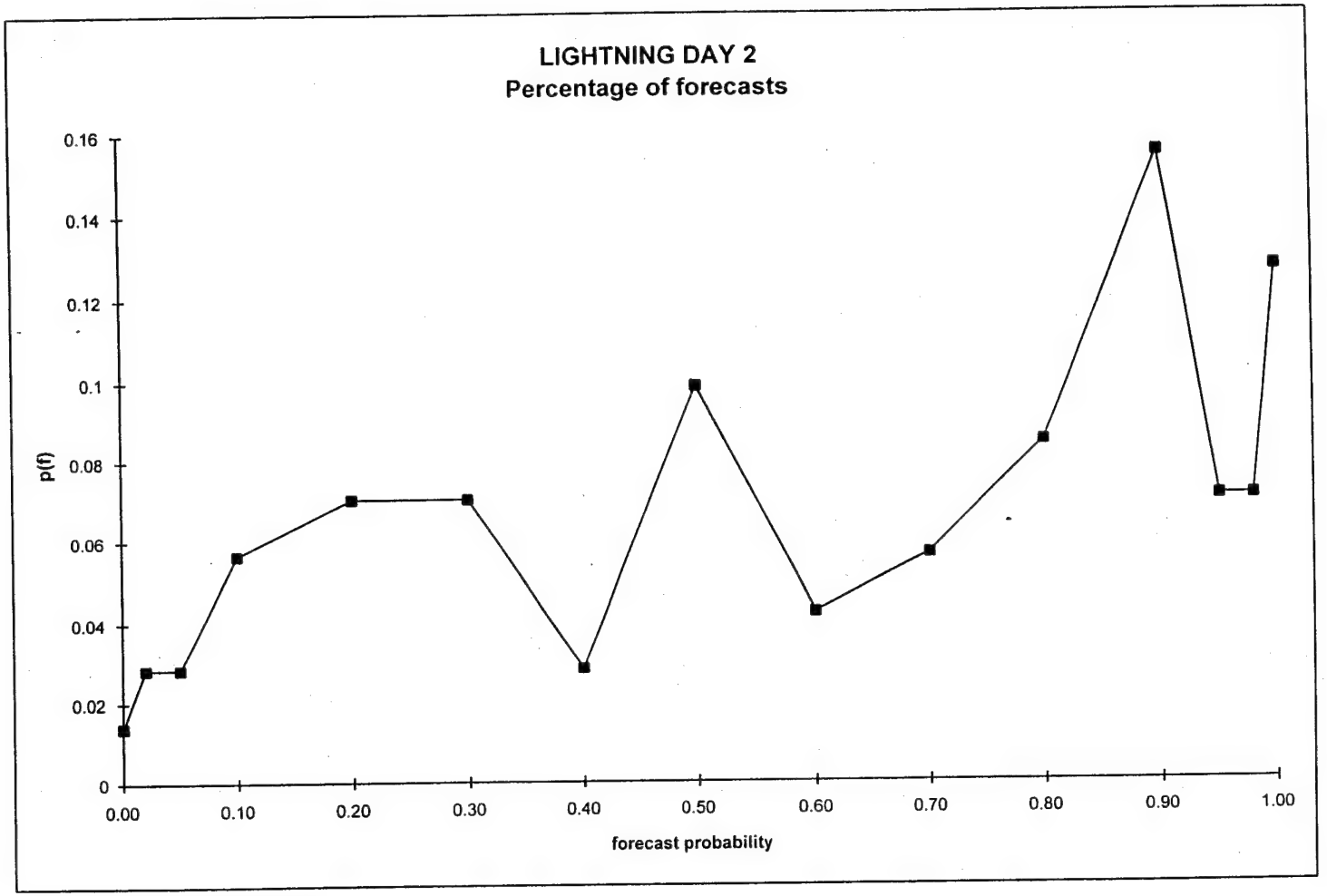


Figure 5a. Percentage of forecasts in each category for lightning Day-2 forecasts.

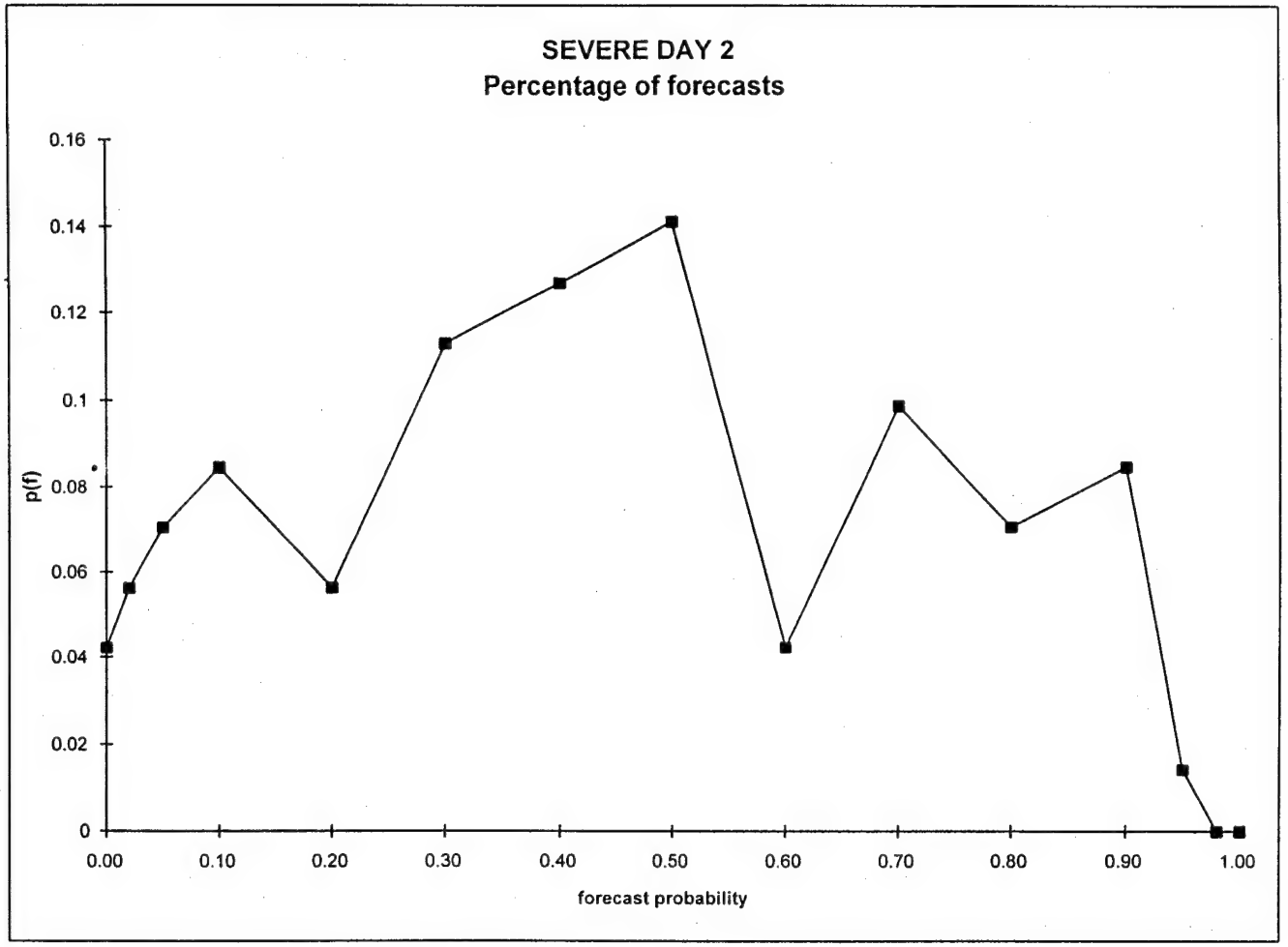


Figure 5b. As in Fig. 5a, except for severe Day-2 forecasts.

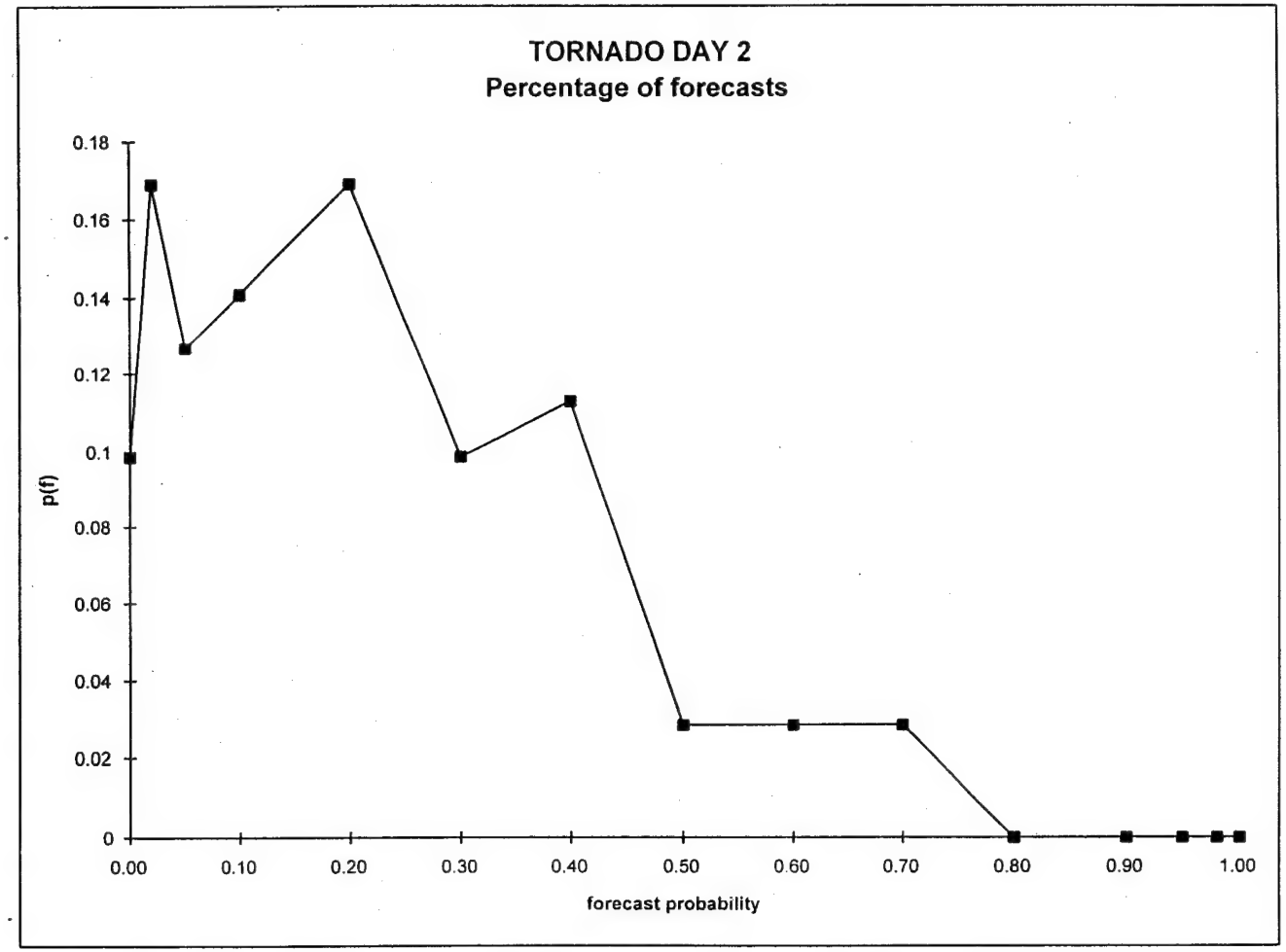


Figure 5c. As in Fig. 5a, except for tornado Day-2 forecasts.

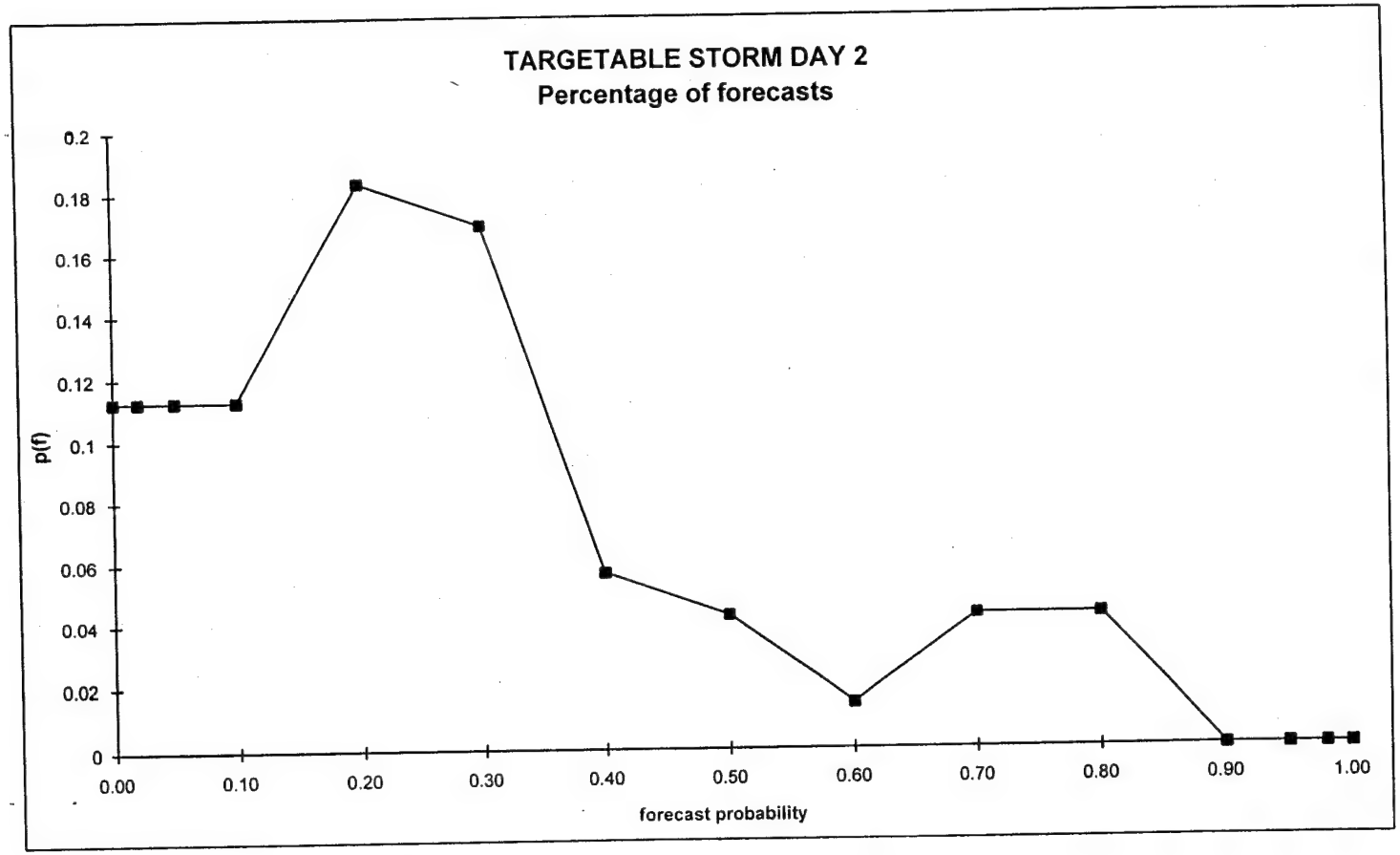


Figure 5d. As in Fig. 5a, except for targetable storm Day-2 forecasts.

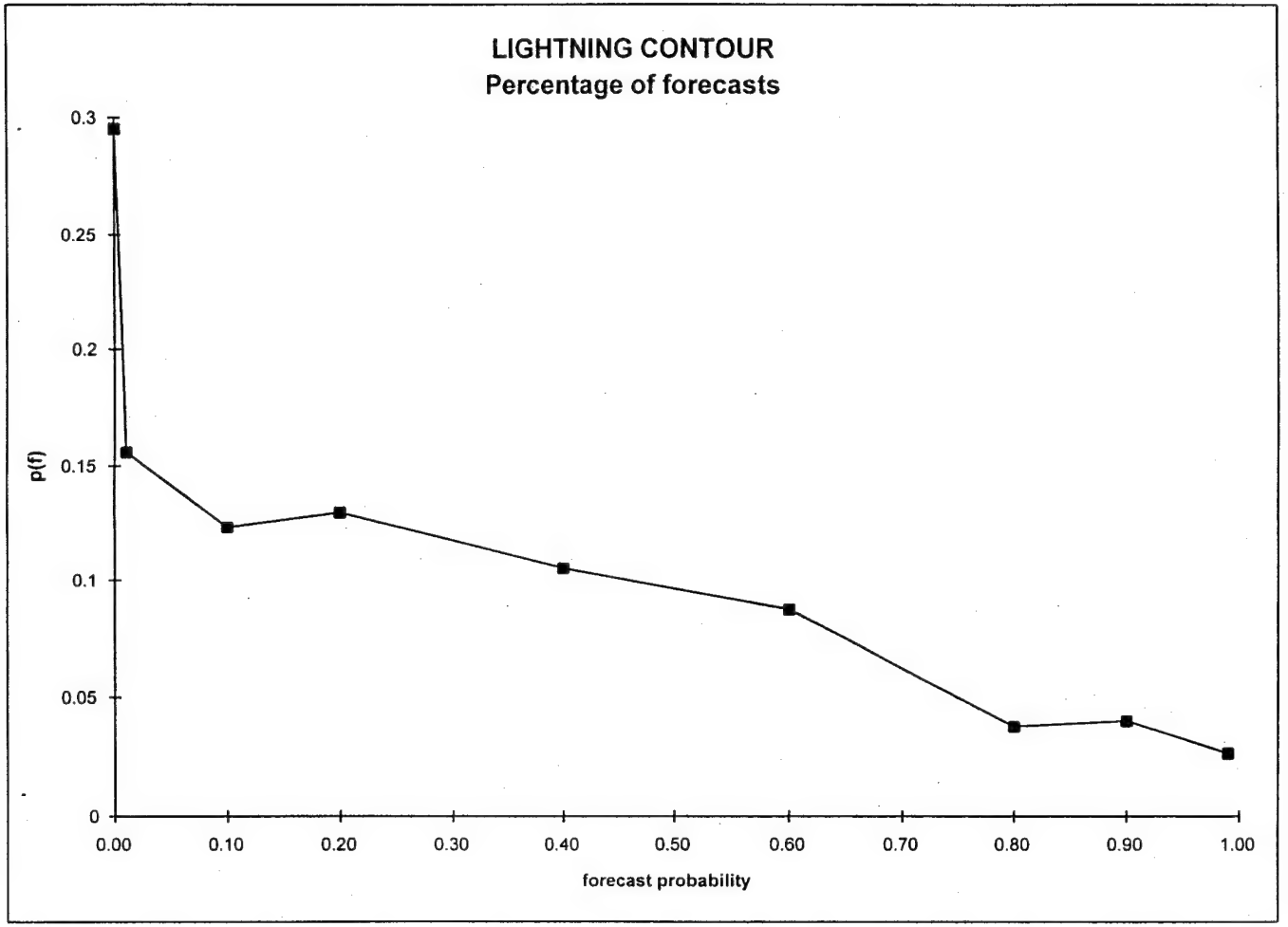


Figure 6a. Percentage of forecasts in each category for lightning contour forecasts.

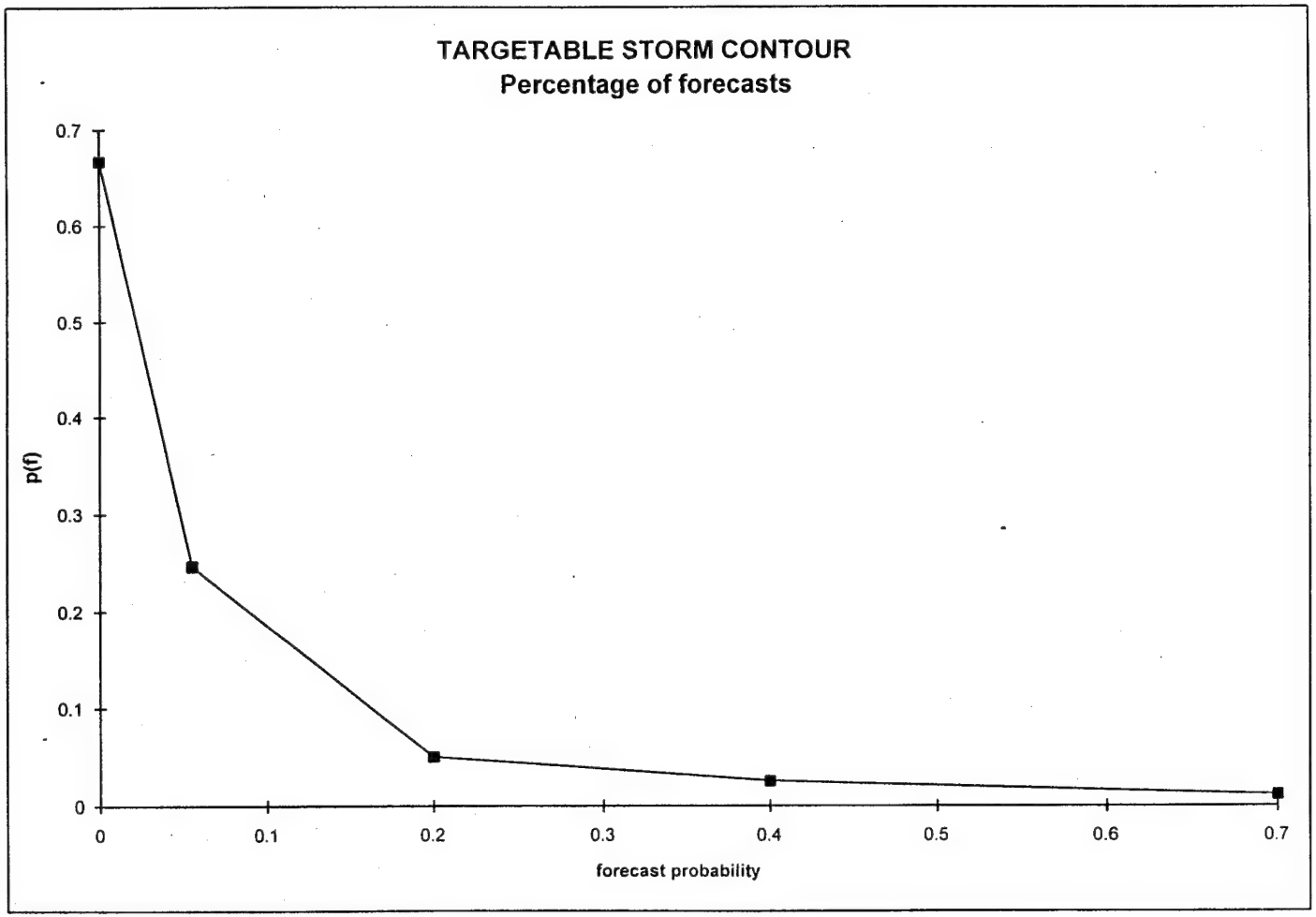


Figure 6b. As in Fig. 6a, except for targetable storm contour forecasts.

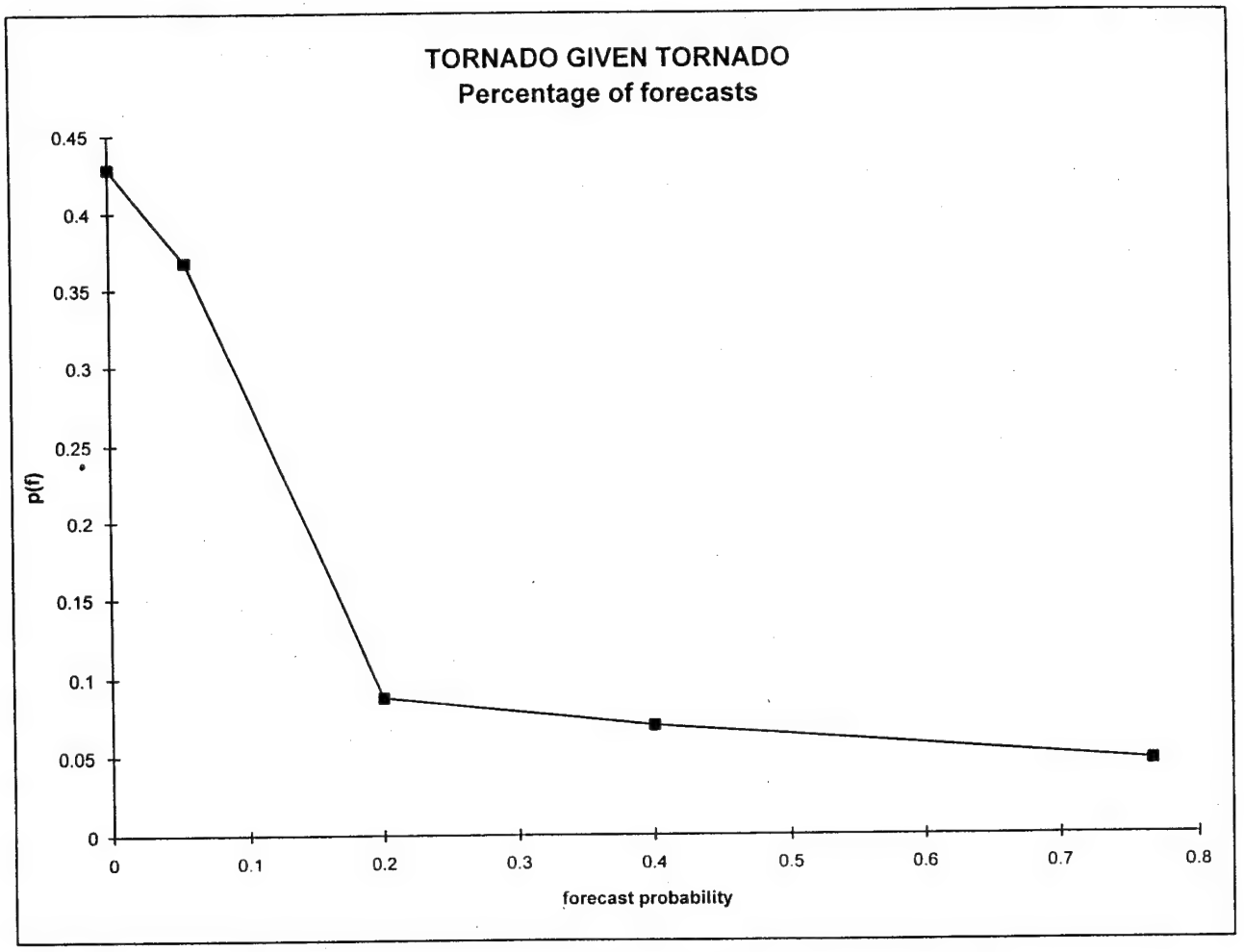


Figure 6c. As in Fig. 6a, except for tornado given tornado contour forecasts.

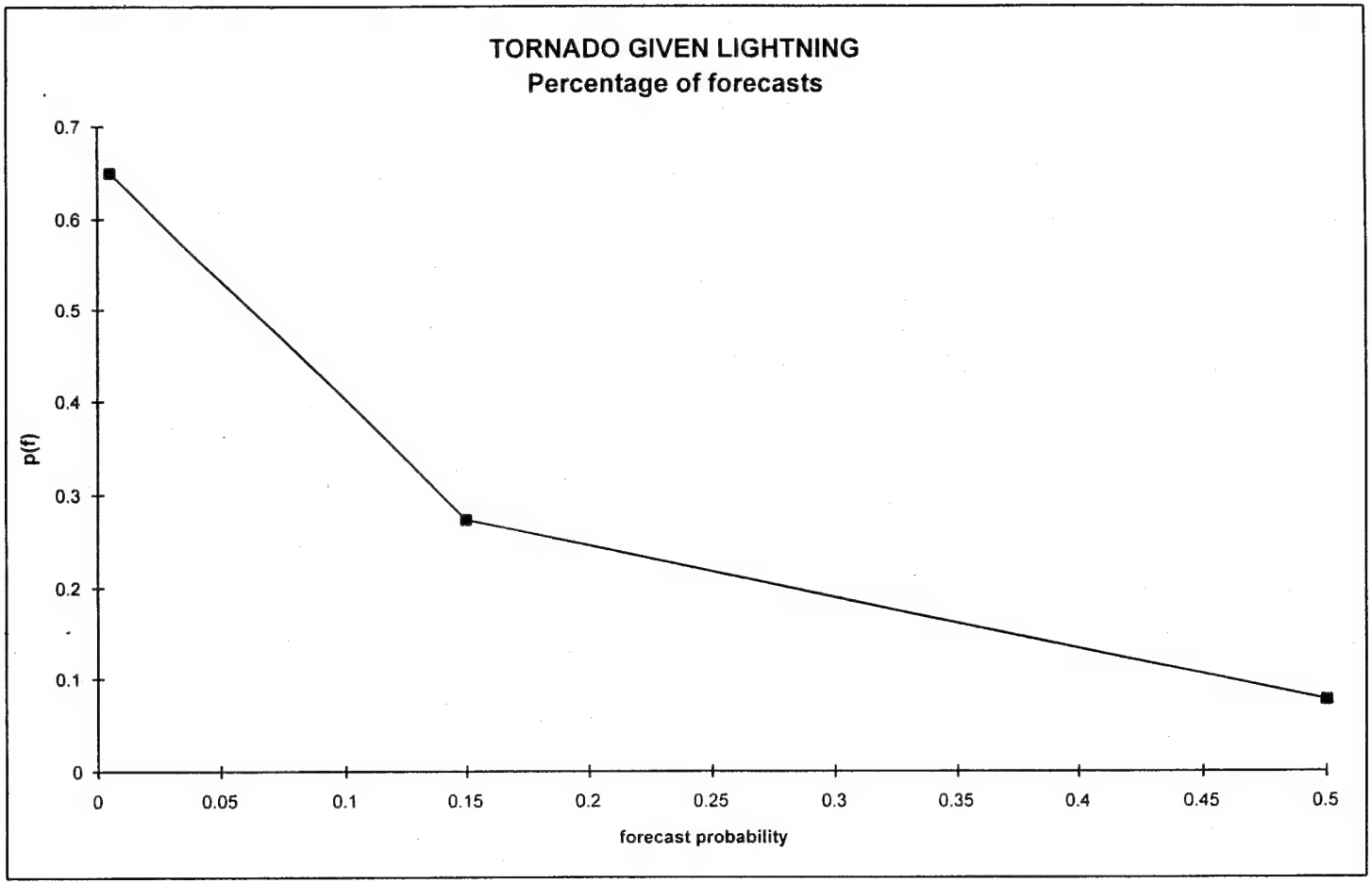


Figure 6d. As in Fig. 6a, except for tornado given lightning contour forecasts.

Table 5a. Conditional probability of an event given the forecast for Day-1 forecasts.

	Lightning		Severe		Tornadoes		Targetable storms	
Fcst	p(x=1|f)	p(x=0|f)	p(x=1|f)	p(x=0|f)	p(x=1|f)	p(x=0|f)	p(x=1|f)	p(x=0|f)
0	0.0	1.0	0.0	1.0	0.0	1.0	.08	.92
2	0.0	1.0	0.0	1.0	.20	.80	0.0	1.0
5	0.0	1.0	.17	.83	.25	.75	.30	.70
10	0.5	0.5	0.5	0.5	0.4	0.6	.56	.44
20	1.0	0.0	0.5	0.5	.22	.78	.38	.62
30	1.0	0.0	.57	.43	.64	.36	0.5	0.5
40	.75	.25	.86	.14	1.0	0.0	.75	.25
50	.88	.12	0.7	0.3	.25	.75	1.0	0.0
60	1.0	0.0	.67	.33	.75	.25	0.5	0.5
70	0.8	0.2	1.0	0.0	1.0	0.0	1.0	0.0
80	1.0	0.0	1.0	0.0	-	-	.83	.17
90	1.0	0.0	1.0	0.0	-	-	-	-
95	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0
98	1.0	0.0	1.0	0.0	-	-	-	-
100	1.0	0.0	1.0	0.0	-	-	-	-

Table 5b. Conditional probability of an event given the forecast for Day-2 forecasts.

	Lightning		Severe		Tornadoes		Targetable storms	
Fcst	p(x=1|f)	p(x=0|f)	p(x=1|f)	p(x=0|f)	p(x=1|f)	p(x=0|f)	p(x=1|f)	p(x=0|f)
0	0.0	1.0	0.0	1.0	0.0	1.0	.13	.87
2	0.0	1.0	0.0	1.0	.25	.75	.13	.87
5	0.5	0.5	0.4	0.6	.22	.78	0.5	0.5
10	0.5	0.5	0.5	0.5	0.3	0.7	0.0	1.0
20	0.6	0.4	0.5	0.5	.58	.42	.54	.46
30	1.0	0.0	.63	.37	.43	.57	0.5	0.5
40	0.5	0.5	.67	.33	.38	.62	0.5	0.5
50	1.0	0.0	0.8	0.2	0.5	0.5	1.0	0.0
60	.67	.33	.67	.33	1.0	0.0	1.0	0.0
70	1.0	0.0	.86	.14	1.0	0.0	1.0	0.0
80	.83	.17	1.0	0.0	-	-	1.0	0.0
90	1.0	0.0	1.0	0.0	-	-	-	-
95	1.0	0.0	1.0	0.0	-	-	-	-
98	1.0	0.0	-	-	-	-	-	-
100	1.0	0.0	-	-	-	-	-	-

Table 5c. Conditional probability of an event given the forecast for lightning and targetable storm contour forecasts.

	L				TS	
Fcst	p(x=1|f)	p(x=0|f)		Fcst	p(x=1|f)	p(x=0|f)
0	.04	.96		0	.00	1.0
1	.05	.95		5.5	.01	.99
10	.14	.86		20	.03	.97
20	.25	.75		40	.06	.94
40	.35	.65		70	.10	.90
60	.39	.61				
80	.46	.54				
90	.66	.34				
99	.79	.21				

Table 5d. Conditional probability of an event given the forecast for conditional tornado probability contour forecasts.

	TGT				TGL	
Fcst	p(x=1|f)	p(x=0|f)		Fcst	p(x=1|f)	p(x=0|f)
0	0.0	1.0		.5	.01	.99
5.5	0.0	1.0		15	.02	.98
20	.02	.98		50	.07	.93
40	.03	.97				
77	.05	.95				

Table 6a. Probability of the forecast given the event for Day-1 forecasts.

	Lightning		Severe		Tornadoes		Targetable storms	
Fcst	p(f|x=1)	p(f|x=0)	p(f|x=1)	p(f|x=0)	p(f|x=1)	p(f|x=0)	p(f|x=1)	p(f|x=0)
0	0.0	.17	0.0	.22	0.0	.20	.03	.25
2	0.0	.17	0.0	.07	.11	.25	0.0	.21
5	0.0	.25	.02	.19	.07	.12	.09	.16
10	.03	.17	.08	.15	.15	.12	.16	.09
20	.05	0.0	.02	.04	.07	.14	.09	.11
30	.02	0.0	.09	.11	.26	.08	.13	.09
40	.05	.08	.12	.04	.04	0.0	.09	.02
50	.11	.08	.14	.11	.04	.06	.09	0.0
60	.03	0.0	.08	.07	.11	.02	.06	.05
70	.06	.08	.10	0.0	.11	0.0	.06	0.0
80	.09	0.0	.08	0.0	0.0	0.0	.16	.02
90	.03	0.0	.10	0.0	0.0	0.0	0.0	0.0
95	.11	0.0	.06	0.0	.04	0.0	.03	0.0
98	.03	0.0	.02	0.0	0.0	0.0	0.0	0.0
100	.39	0.0	.08	0.0	0.0	0.0	0.0	0.0

Table 6b. Probability of the forecast given the event for Day-2 forecasts.

	Lightning		Severe		Tornadoes		Targetable storms	
Fcst	p(f|x=1)	p(f|x=0)	p(f|x=1)	p(f|x=0)	p(f|x=1)	p(f|x=0)	p(f|x=1)	p(f|x=0)
0	0.0	.09	0.0	.12	0.0	.16	.03	.18
2	0.0	.18	0.0	.16	.12	.20	.03	.18
5	.02	.09	.04	.12	.08	.16	.13	.10
10	.03	.18	.07	.12	.12	.16	0.0	.20
20	.05	.18	.04	.08	.27	.11	.23	.15
30	.08	0.0	.11	.12	.12	.09	.19	.15
40	.02	.09	.13	.12	.12	.11	.07	.05
50	.12	0.0	.17	.08	.04	.02	.10	0.0
60	.03	.09	.04	.04	.08	0.0	.03	0.0
70	.07	0.0	.13	.04	.08	0.0	.10	0.0
80	.08	.09	.11	0.0	0.0	0.0	.10	0.0
90	.18	0.0	.13	0.0	0.0	0.0	0.0	0.0
95	.08	0.0	.02	0.0	0.0	0.0	0.0	0.0
98	.08	0.0	0.0	0.0	0.0	0.0	0.0	0.0
100	.15	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 6c. Probability of the forecast given the event for lightning and targetable storm contour forecasts.

	L				TS	
Fcst	p(f |x=1)	p(f |x=0)		Fcst	p(f |x=1)	p(f |x=0)
0	.06	.36		0	.11	.67
1	.04	.19		5.5	.26	.25
10	.09	.13		20	.23	.05
20	.16	.12		40	.23	.02
40	.18	.09		70	.17	.01
60	.17	.07				
80	.08	.03				
90	.13	.02				
99	.10	.01				

Table 6d. Probability of the forecast given the event for conditional tornado probability contour forecasts.

	TGT				TGL	
Fcst	p(f |x=1)	p(f |x=0)		Fcst	p(f |x=1)	p(f |x=0)
0	.13	.43		.5	.26	.66
5.5	.17	.37		15	.38	.27
20	.17	.09		50	.36	.07
40	.28	.07				
77	.26	.05				

4.5.1 Summary measures

The reliability, resolution, and discrimination of the forecasts can be reduced to summary measures, giving a first look at the overall quality of the forecasts as measured by these indicators. Although these numbers offer some insight, they must be interpreted with caution, as with any summary measure. The **reliability** (or calibration) looks at the difference between the conditional mean observation and the conditioning forecast, averaged over all forecasts. This is calculated using the following formula (Murphy 1993, Wilks 1995):

$$\text{Reliability} = \frac{\sum_i D_i (P(X = 1|F_i) - F_i)^2}{n}. \qquad (16)$$

This number should be close to zero for reliable forecasts, since the squared differences between the conditional mean forecast and the conditioning forecast should be minimized.

The **resolution** indicates the ability of the forecasts to resolve between days that are more or less likely than climatology to observe an event. It is defined as the difference between the conditional mean observation and the unconditional mean observation, averaged over all forecasts (Murphy 1993, Wilks 1995),

$$\text{Resolution} = \frac{\sum_i D_i (P(X = 1|F_i) - \overline{X})^2}{n}. \qquad (17)$$

Large differences between the conditional mean observation and the unconditional mean observation indicate that the forecasts exhibit good resolution. If the conditional mean observation is close to the sample climatology, then the observations are approximately independent of the forecasts.

The **discrimination** relates to the ability of the forecasts to discriminate or distinguish among the observations. It has two components: the first type of discrimination is given as the difference between the conditional mean forecast and the conditioning observation, averaged over all observations (Murphy 1993),

$$\text{Discrimination 1} = \frac{H\left(\sum_i F_i\left(P(F_i|X=1)\right)-1\right)^2 + N\left(\sum_i F_i\left(P(F_i|X=0)\right)-0\right)^2}{n}. \quad (18)$$

The conditional mean forecast given a hit (non-event) should be close to one (zero) for small differences in both terms in the numerator and, hence, good discrimination. The second type of discrimination is defined by the difference between the conditional mean forecast and the unconditional mean forecast, averaged over all observations (Murphy 1993),

$$\text{Discrimination 2} = \frac{H\left(\sum_i F_i\left(P(F_i|X=1)\right)-\bar{F}\right)^2 + N\left(\sum_i F_i\left(P(F_i|X=0)\right)-\bar{F}\right)^2}{n}. \quad (19)$$

The first term is the difference between the conditional mean forecast given an event and the unconditional mean forecast, while the second term is the difference between the conditional mean forecast given a non-event and the unconditional mean forecast. Large differences are preferred. If there were no differences between the conditional mean forecasts and the unconditional mean forecasts, the forecasts would be independent of the observations.

4.5.2 Reliability, Resolution, and Discrimination - numerical

These summary measures are presented numerically in Tables 7a-c. Beneath each value is the range of possible values for each measure. The best (worst) possible value is obtained by assuming all hits occur with a forecast of 100% (0%) and all non-events occur with a forecast of 0% (100%). Recall that low (high) values are desired for reliability (resolution) and low (high) values are also desired for discrimination 1 (discrimination 2). The best forecast for each quality indicator is shown in **bold** letters.

For Day-1, the reliability was best (worst) for tornado (lightning) forecasts, whereas the resolution for CG lightning is slightly better than for tornado forecasts. This substantiates our earlier assertion that a forecast with good reliability does not necessarily show good resolution. In this case, the tornado forecasts were closer to climatology and, thus, showed good reliability but poor resolution. For comparison purposes, reliability values obtained in past studies were, for: probabilistic QPF forecasts (.026-.035) (Murphy et al. 1985); PoP forecasts in a field experiment (.056) (Murphy and Daan 1984), probabilistic tornado forecasts within SELS outlook and watch areas (.004 and .026, respectively) (Murphy and Winkler 1982). These lower reliability values indicate that these forecasts were apparently more reliable than those issued during VORTEX '94 but direct comparison is difficult because of the many

Table 7a. Reliability, Resolution, and Discrimination values by phenomenon for Day-1.

Measure	L	S	T	TS
reliability	.070 (0-1)	.061 (0-1)	**.052** **(0-1)**	.061 (0-1)
resolution	.088 (0-.37)	**.106** **(0-.27)**	.072 (0-.27)	.090 (0-.26)
discrimination 1	**.045** **(0-1)**	.117 (0-1)	.170 (0-1)	.162 (0-1)
discrimination 2	**.050** **(0-.28)**	.039 (0-.25)	.010 (0-.35)	.018 (0-.32)

Table 7b. As in Table 7a, for Day-2.

Measure	L	S	T	TS
reliability	.093 (0-1)	.067 (0-1)	**.052** **(0-1)**	.074 (0-1)
resolution	.065 (0-.37)	.077 (0-.27)	.050 (0-.27)	**.091** **(0-.25)**
discrimination 1	**.077** **(0-1)**	.158 (0-1)	.205 (0-1)	.194 (0-1)
discrimination 2	**.031** **(0-.27)**	.023 (0-.26)	.005 (0-.35)	.012 (0-.33)

Table 7c. As in Table 7a, for contour forecasts.

Measure	L	TS	TGT	TGL
reliability	.013 (0-1)	**.008** **(0-1)**	.038 (0-1)	.019 (0-1)
resolution	**.037** **(0-.33)**	.000 (0-.49)	.000 (0-.49)	.000 (0-.49)
discrimination 1	.072 (0-1)	**.005** **(0-1)**	.013 (0-1)	.014 (0-1)
discrimination 2	**.018** **(0-.32)**	.000 (0-.46)	.001 (0-.41)	.000 (0-.43)

differences in the forecasting and verification methods used between the different studies.

The resolution of the forecasts is related to skill, which is why the severe (Day-1) forecasts showed the best resolution. For comparison purposes, resolution values obtained in past studies were, for: probabilistic QPF forecasts (.049-.088) (Murphy et al. 1985); PoP forecasts in a field experiment (.021) (Murphy and Daan 1984); probabilistic tornado forecasts within SELS outlook and watch areas (.05 and .02, respectively) (Murphy and Winkler 1982). VORTEX '94 forecasts do show better resolution than these past studies, but again, direct comparison is not possible.

The discrimination is best for lightning forecasts and worst for tornado forecasts. This is also the result of forecasts that are closer to climatology, since the mean tornado forecast, given a hit, is farther from one and closer to the unconditional mean forecast than for other forecast phenomena. Discrimination values were not computed in the past studies examined.

For contour forecasts, the results are dominated by the large number of correctly forecast non-events. These summary measures indicate that targetable storms are the most reliable forecasts, but it will become obvious from the distributions (presented later) that this is definitely not the case. In the same way that summary measures like the Brier score were unreliable for rare events, these summary measures are also unreliable. When these are presented graphically, it's clear that lightning forecasts have the best reliability, resolution, and discrimination.

Looking at the these summary measures numerically doesn't add much to our understanding of the relationships between forecasts and observations. These measures point us in the right direction for which forecasts have the best overall reliability, resolution, and discrimination but they still don't indicate **where** in the

forecast range the forecasts are particularly good or bad. To understand these relationships properly requires that these quality indicators be examined graphically.

4.5.3 Reliability, Resolution, and Discrimination - graphical

4.5.3.1 Lightning area forecasts

The **reliability** diagram for the lightning area forecasts is presented in Fig. 7a. Note that lightning forecasts were the least reliable. The goal is to have a reliability curve close to the perfect reliability line, or at least to have the observed frequency increasing monotonically as the forecast probability increases. It's important to keep in mind that the peaks and valleys on these diagrams sometimes only represent 2 or 3 forecasts, so they may not be statistically significant. For the graphical presentation, it's more important to look at the trend in the curves rather than the individual peaks and valleys. One of the values of using a plot like this is noting specifically **where** the underforecasting is occurring. This plot shows a lot of underforecasting above a forecast probability of 10%. The forecasters were apparently not confident enough to use higher probabilities, though their use was warranted. They are underestimating the frequency with which lightning (and all other phenomena) occur in this relatively large area. There is some evidence to suggest that the areal coverage was at least smaller on those days that smaller probability values were used. The five hits that occurred below 30%, for example, only exhibited an areal coverage of about 7% of the area, compared to about 24% of the area hit on an average lightning day. This

verification made no distinction between those days with a few hits and those days with 1,000 hits. This may explain some of the underforecasting. One way to account for the relative number of events would be to verify observed areal coverage (e.g., the percentage of MDR grid boxes hit). Improving the reliability of the forecasts means reducing these biases.

The **resolution** diagram for lightning forecasts is shown in Fig. 7b. Better resolution is obtained if all events occur when the forecast is above climatology, and all non-events occur when the forecast is below climatology. This would have all forecasts above climatology with an observed frequency of one, and all forecasts below climatology with an observed frequency of zero (a very skillful forecast but not necessarily reliable (e.g. could be biased)). This means that the distance of the $p(x=1 \mid f)$ line above (below) the climatological frequency line should be maximized when the forecast is above (below) climatology for the best possible resolution. The CG lightning forecasts show good resolution for forecasts above climatology and for forecasts at very low probabilities but there is a large range in the middle of the forecast range (where there was also underforecasting) where the resolution is poor. The forecasts in this range are not skillful. The resolution for lightning forecasts was slightly better than tornado forecasts, however. Improving the resolution of the forecasts requires, first, knowing the climatology of the events being forecast, then, forecasting as far above climatology as possible (based on forecaster confidence) when

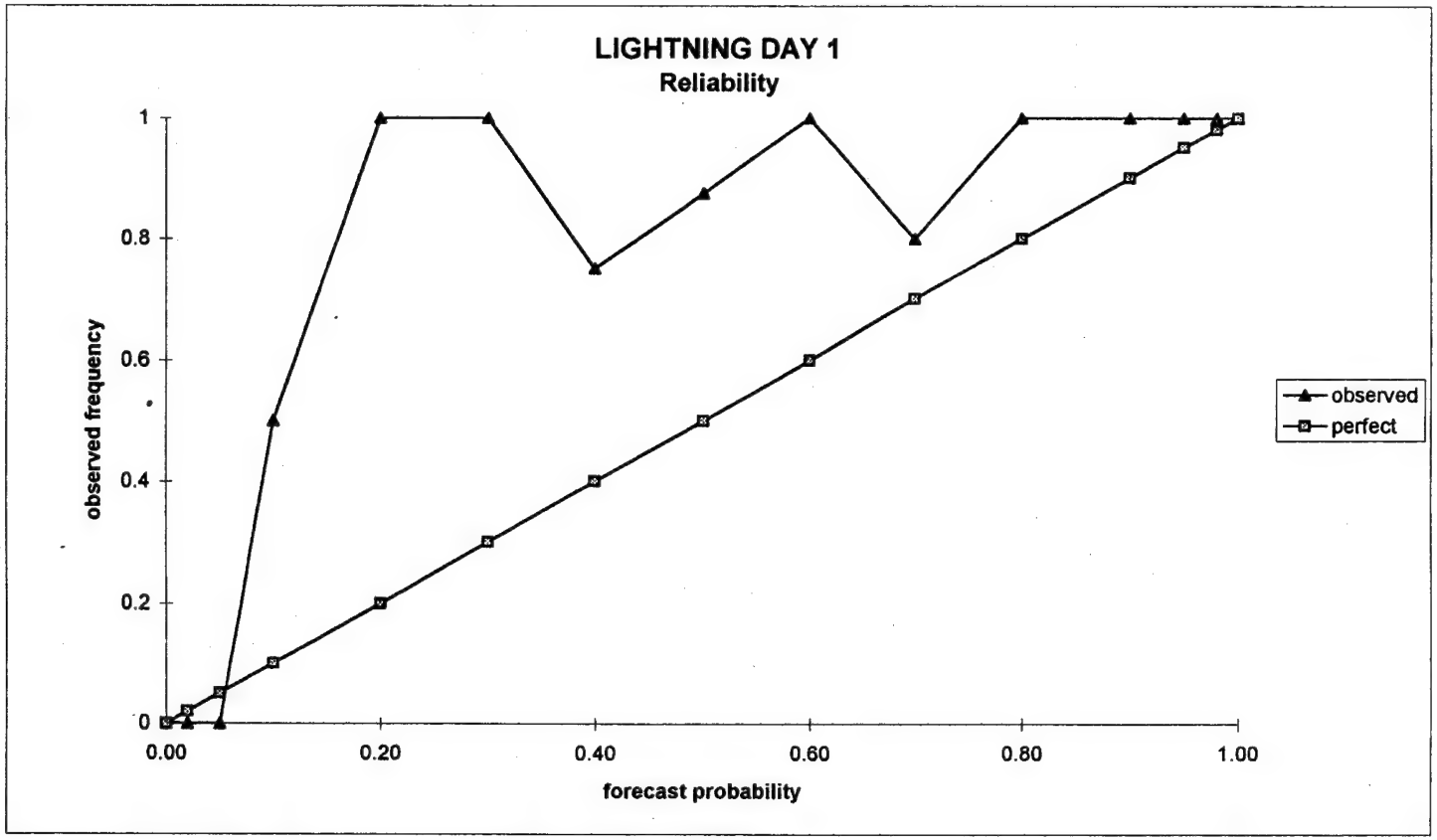


Figure 7a. Reliability diagram for lightning Day-1 forecasts.

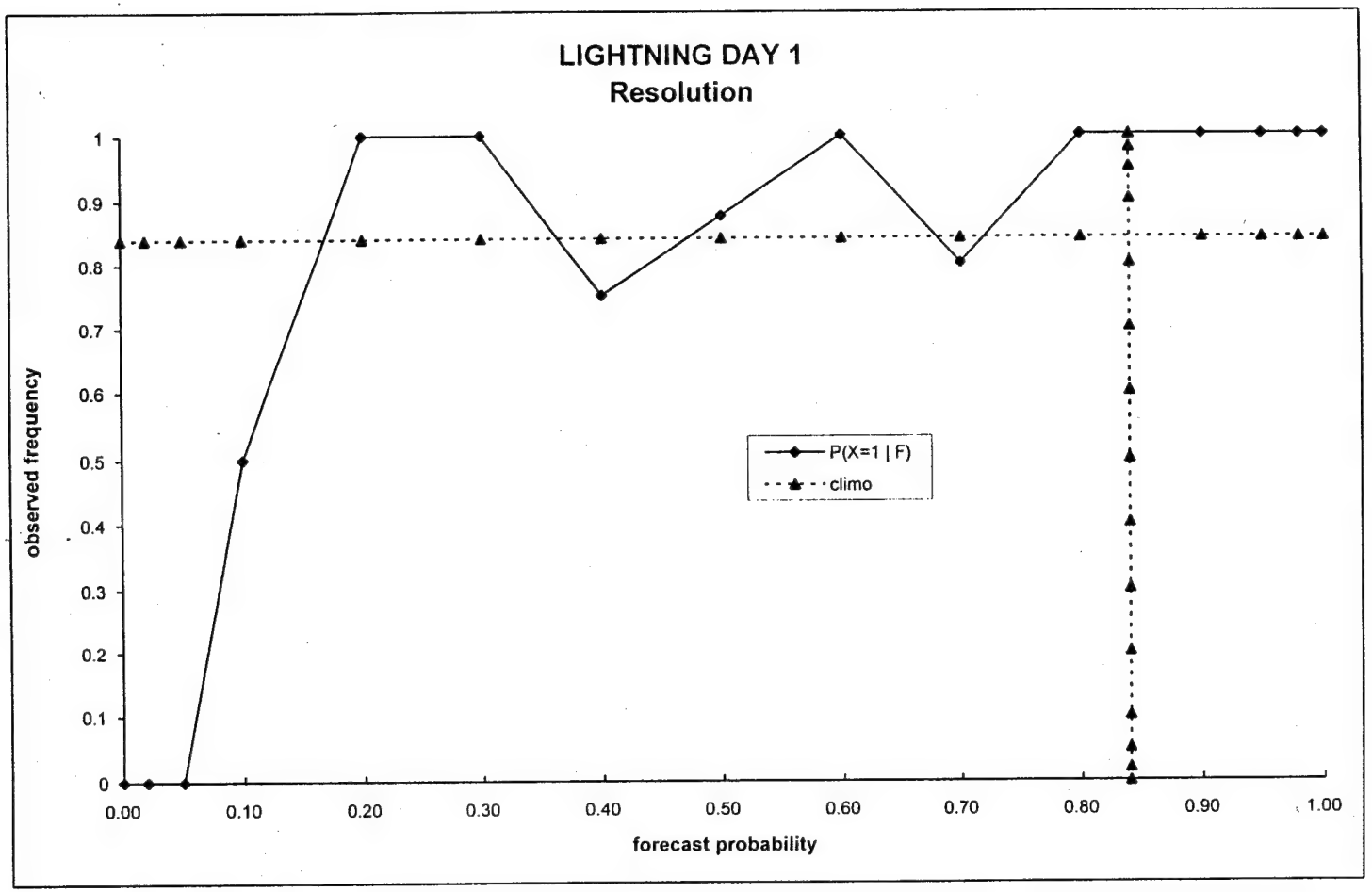


Figure 7b. Resolution diagram for lightning Day-1 forecasts.

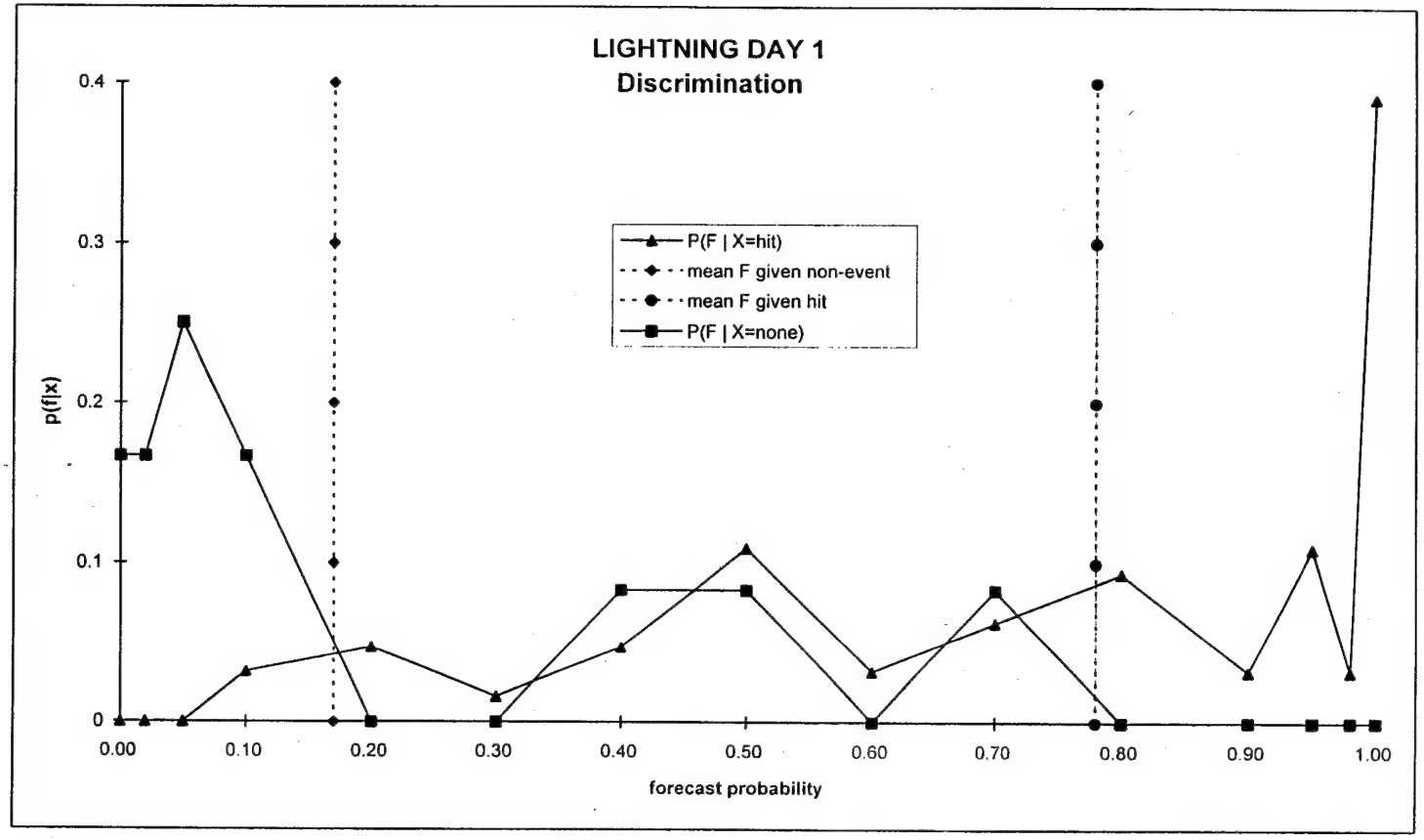


Figure 7c. Discrimination diagram for lightning Day-1 forecasts.

an event is expected and as far below climatology as possible when a non-event is expected.

The CG lightning forecasts showed the best **discrimination** (Fig. 7c), with the mean forecasts being closer to their respective ideal values (zero and unity) than for any other forecast. For the lightning forecasts, the large number of of correct forecasts at 100% helps the discrimination considerably. There is a range below 100%, however, where the $p(f \mid x=1)$ curve has a flatter distribution, indicating less discrimination. Discrimination would also improve by having forecasts as close to one as possible when an event is expected, and as close to zero as possible when a non-event is expected.

4.5.3.2 Tornado area forecasts

In the tornado **reliability** diagram (Fig. 8a), the trend is for the curve to be closer to the perfect reliability line and, indeed, this was the most reliable forecast. Recall that the individual peaks and valleys may not be statistically significant. There is a larger number of correct forecasts at the low end of the probability scale, as expected, which helps the reliability number (computed earlier) considerably. The lightning and tornado forecasts both show approximately equal amounts of underforecasting. The key to which one had better reliability was identifying where this underforecasting was occurring. Having hits occur in the lower probabilities was more damaging to lightning forecasts than tornado forecasts because of the higher

relative frequency of lightning. Reducing the biases would improve the reliability even further.

The **resolution** diagram for tornado forecasts is shown in Fig. 8b. Though the resolution appears to be good in the higher probabilities, there are fewer forecasts in these categories than in the middle of the forecast range where the resolution is not so good. It is for this reason that the tornado forecasts showed the worst resolution. The forecasts with the best resolution (severe weather), also had the highest skill scores (20%), while those with the worst resolution (tornadoes), had the lowest skill scores (9%). This makes sense, since forecasts that stick closer to climatology will show less of an improvement over climatology. Whereas the skill score can tell us which forecasts improved over climatology the most, the resolution diagrams can tell us exactly where the skill of the forecasts lies.

The **discrimination** diagram for tornado forecasts is shown in Fig. 8c. Tornado forecasts showed the poorest discrimination primarily because of the poor discrimination on the upper end of the forecast scale. There is some indication of discrimination, however, in the 0-30% range. Since this is where the majority of the forecasts fell, this is somewhat encouraging. The results of this study suggest that the discrimination decreases as the events become rarer. Rare event forecasts typically use lower probabilities, which results in less discrimination in the higher probabilities.

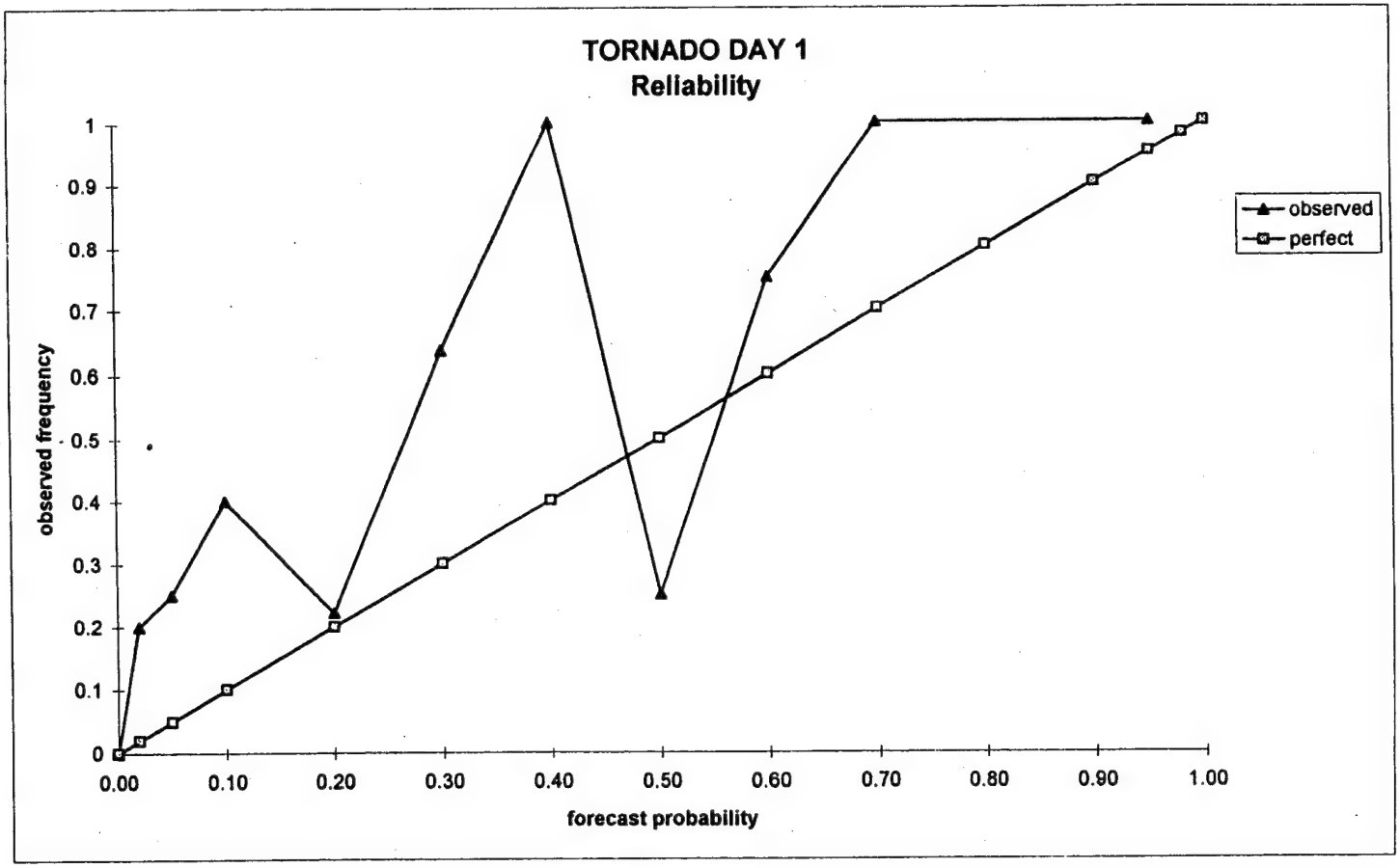


Figure 8a. Reliability diagram for tornado Day-1 forecasts.

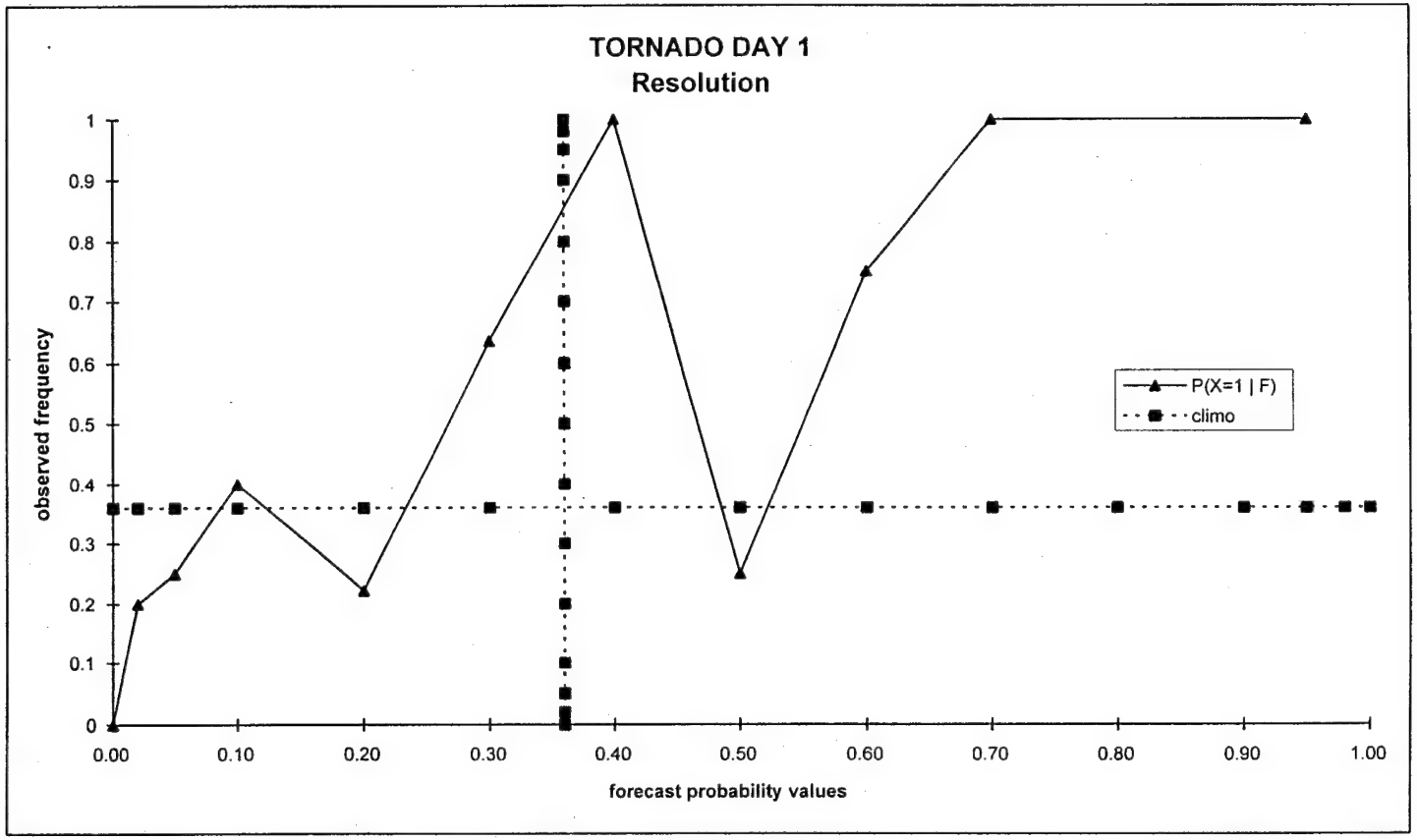


Figure 8b. Resolution diagram for tornado Day-1 forecasts.

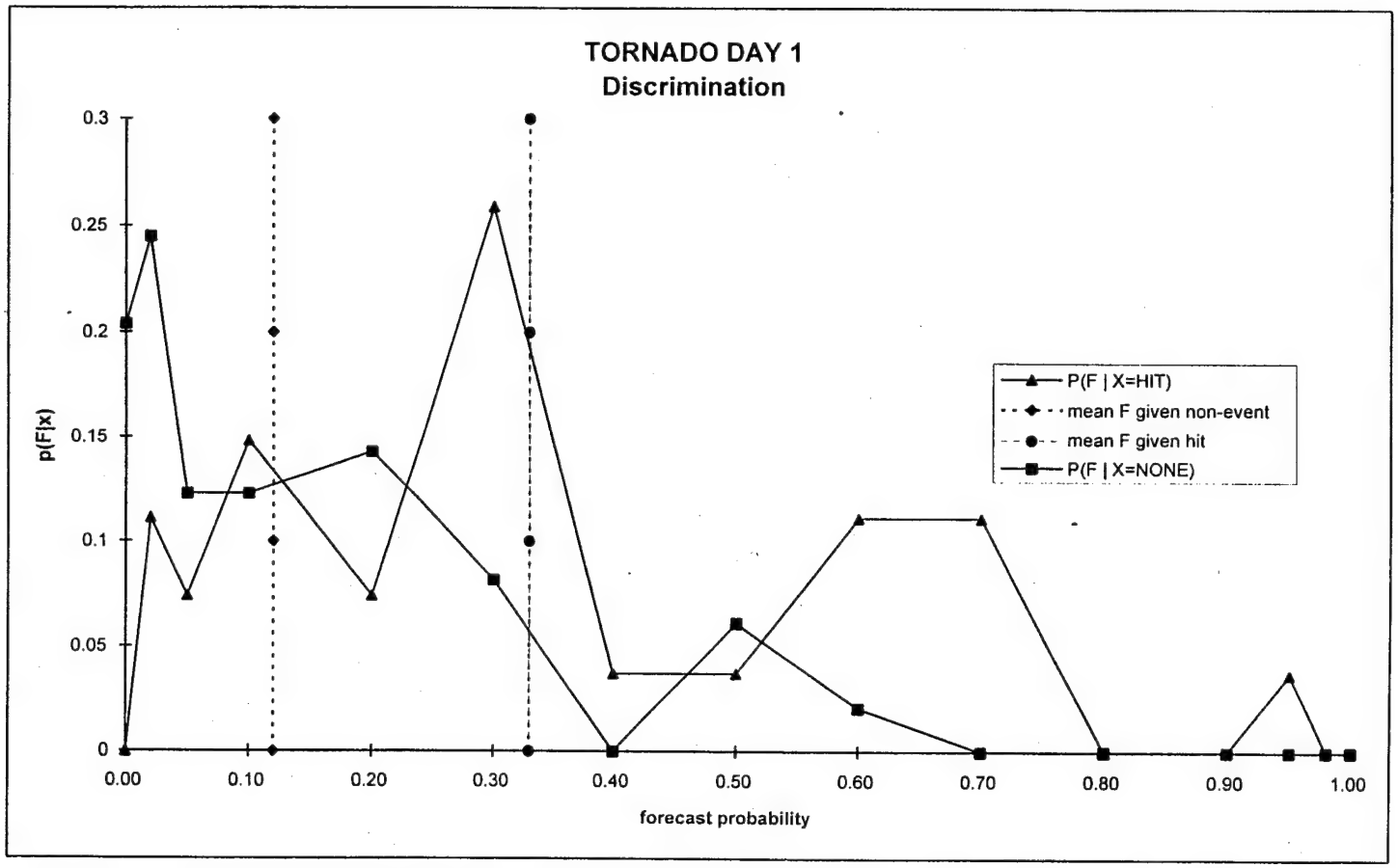


Figure 8c. Discrimination diagram for tornado Day-1 forecasts.

4.5.3.3 Lightning contour forecasts

For contour forecasts, the percentage of each probability area hit is the important factor in determining the **reliability** of the forecast. The CG lightning forecasts show very good reliability (Fig. 9a) in the lower probabilities, with overforecasting occurring in the higher probabilities. The CG lightning forecasts were the most reliable of the contour forecasts.

The **resolution** of the CG lightning contour forecasts is shown in Fig. 9b. CG lightning contour forecasts indicate considerably more skill than the other contour forecasts, and this skill is present over the entire forecast range. This suggests that the forecasters do show some ability in pinpointing the location of convection.

The **discrimination** is also better for CG lightning (Fig. 9c) than for any other contour forecast. There is a small degree of discrimination below 40%, with little if any discrimination above this. The combination of these quality indicators all indicate that forecasts above 40% were not good. The forecasters are apparently overly confident in their ability to identify the location of deep convection. In light of this information, they should weigh carefully the decision to use probabilities greater than 40%.

4.5.3.4 Conditional probability TGT

The results for targetable storm and conditional tornado probability forecasts all show similar results for reliability, resolution, and discrimination so we'll only look at the TGT forecasts.

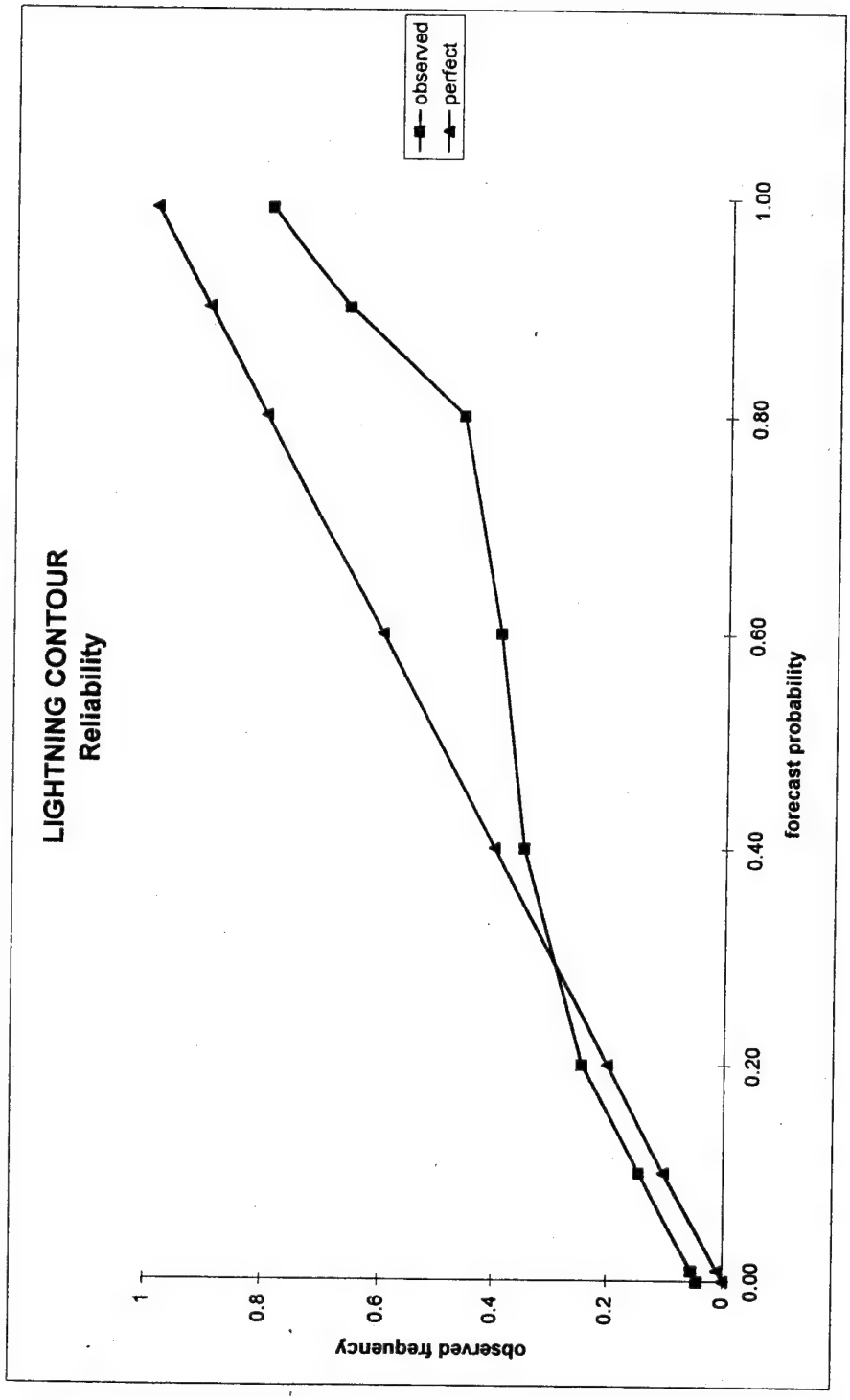


Figure 9a. Reliability diagram for lightning contour forecasts.

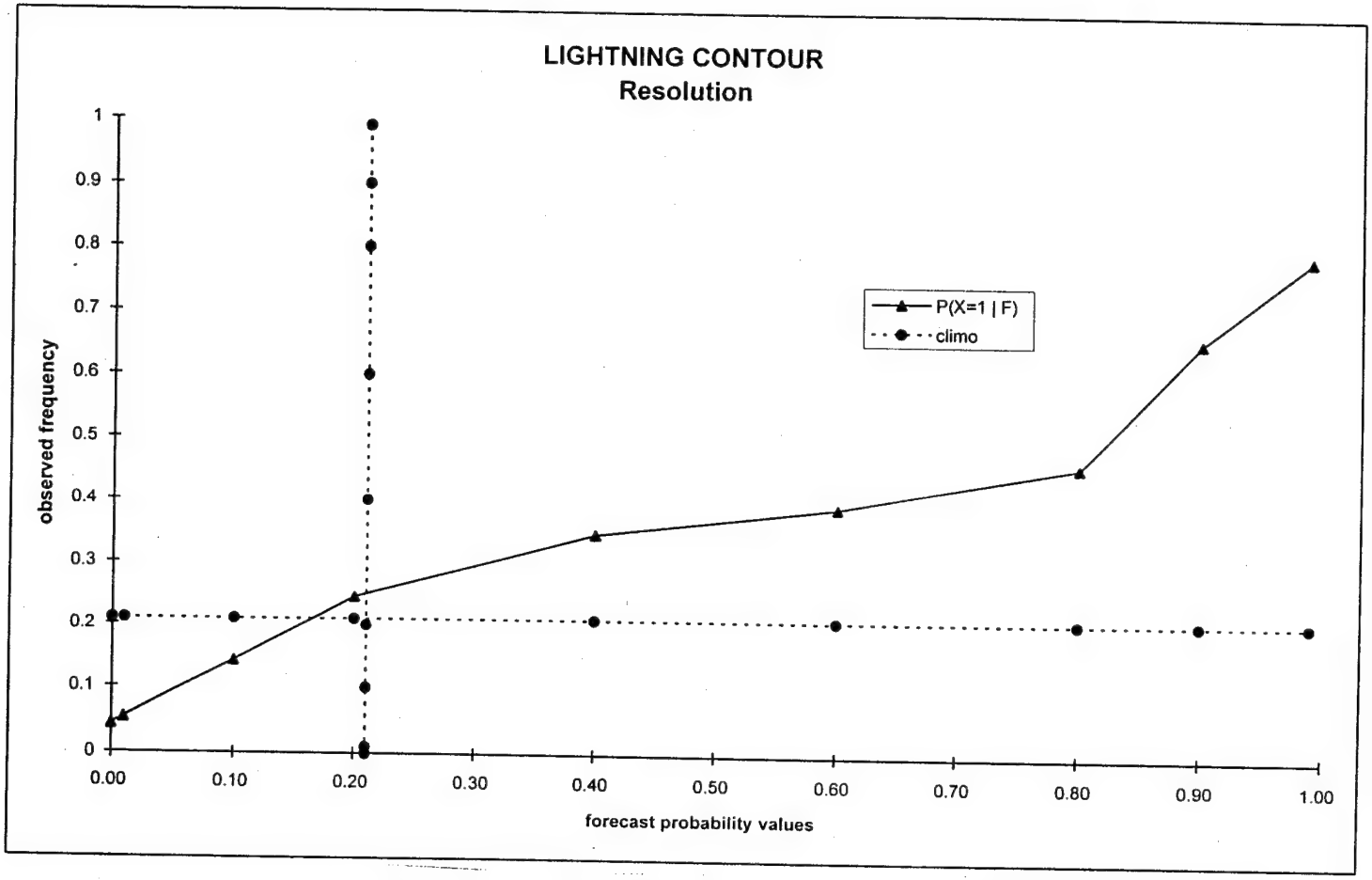


Figure 9b. Resolution diagram for lightning contour forecasts.

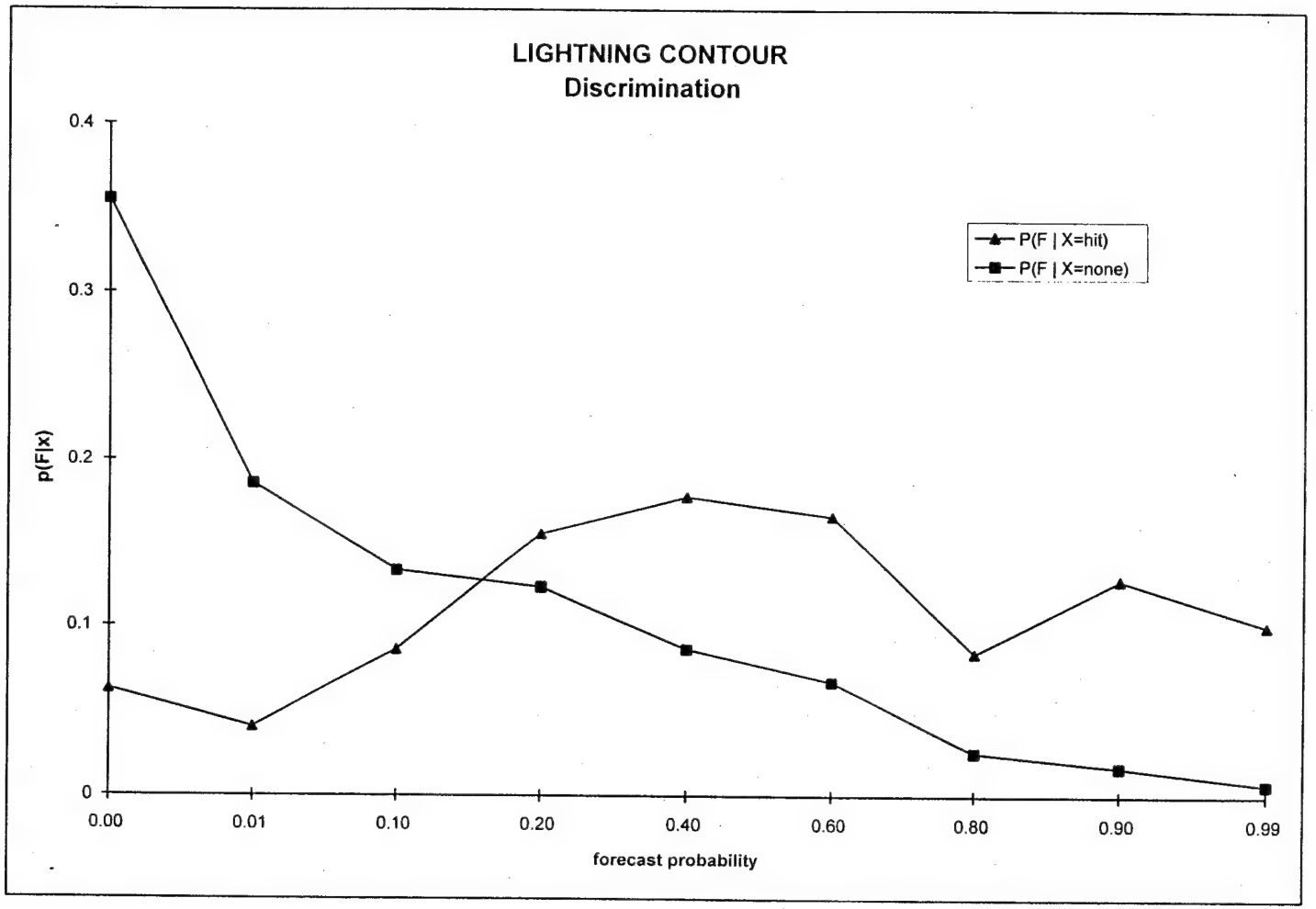


Figure 9c. Discrimination diagram for lightning contour forecasts.

The **reliability** (Fig. 10a) is very poor, with very low event frequencies over the entire forecast range resulting in a lot of overforecasting. This is caused by the small spatial coverage of the observed events (their rarity) compared to the much larger area covered by the probability contours. In general, the forecasters used probabilities that were too high over areas that were too large for an MDR-size grid. It's quite possible that it was asking too much of our forecasters to provide this much spatial specificity. Therefore, the verification of these contour forecasts was re-done with decreased spatial resolution, using larger grid sizes (2 X 2, 3 X 3, 4 X 4, and 5 X 5 MDR boxes). The probabilities forecast in each MDR box were averaged to come up with a forecast for the larger boxes. The results are shown in Figs. 11a-d, where it's obvious that the reliability depends on the grid size. This is only an experiment, however, since the forecasters might have forecast different probabilities if they had known the grid size would be larger.

Not surprisingly, the targetable storm and conditional tornado contour forecasts show very little **resolution**. The resolution diagram for TGT contour forecasts is shown in Fig. 10b. The forecasts strayed very little from climatology, but the conditional probability of an event given the forecast is at least higher than the climatological probability. This says that the forecasters have at least shown **some** ability at identifying days that are more likely than climatology to observe an event. It's encouraging that at least the observed frequency of events (non-events) is increasing (decreasing) with increasing forecast probability, though only slightly.

The tornado given tornado forecasts also show little **discrimination** (Fig. 10c). There is some small degree of discrimination for forecasts below 40% but little if any discrimination above this. This says that given an event, a forecast of 40% is more likely than any of the lower probabilities.

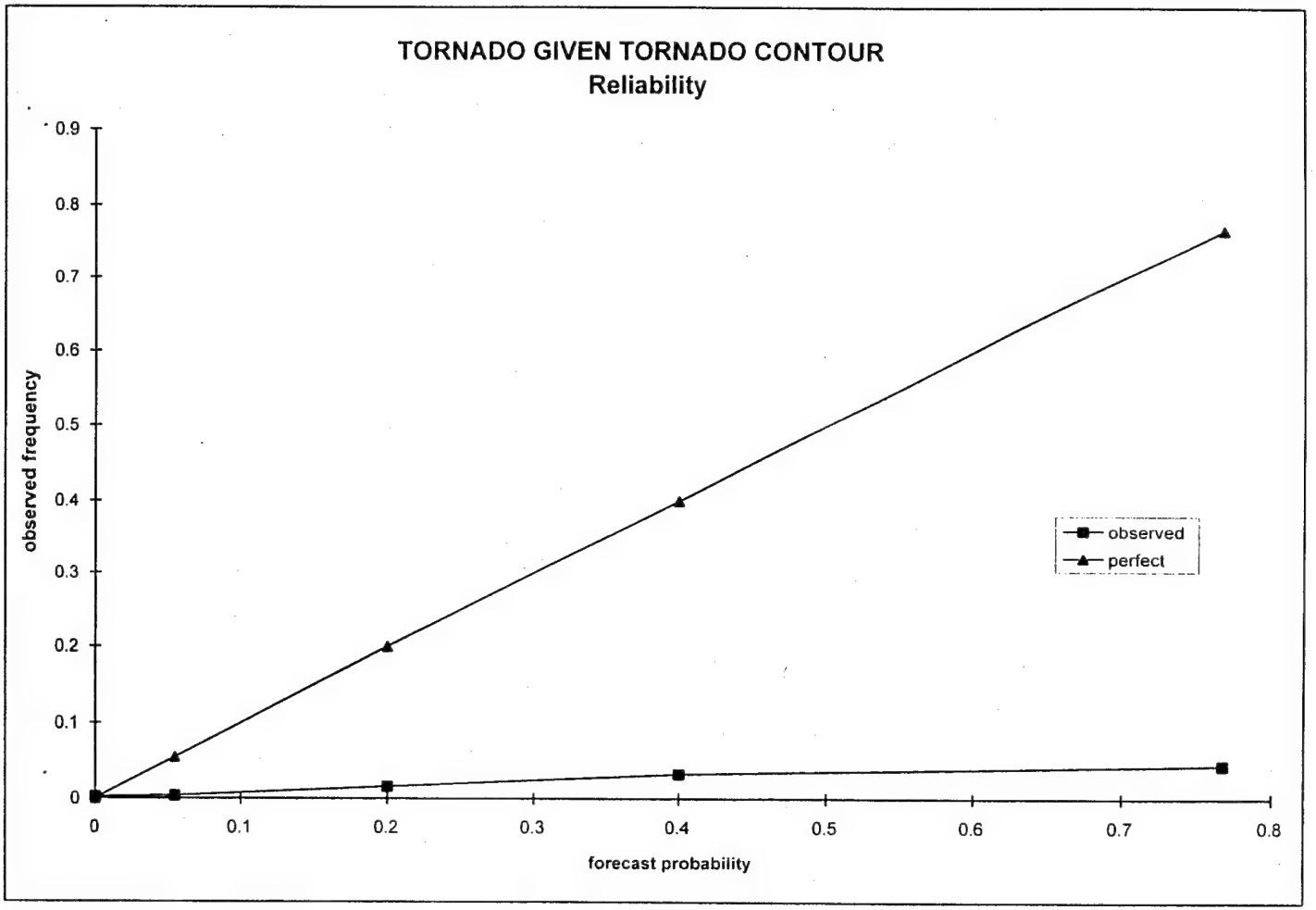


Figure 10a. Reliability diagram for tornado given tornado contour forecasts.

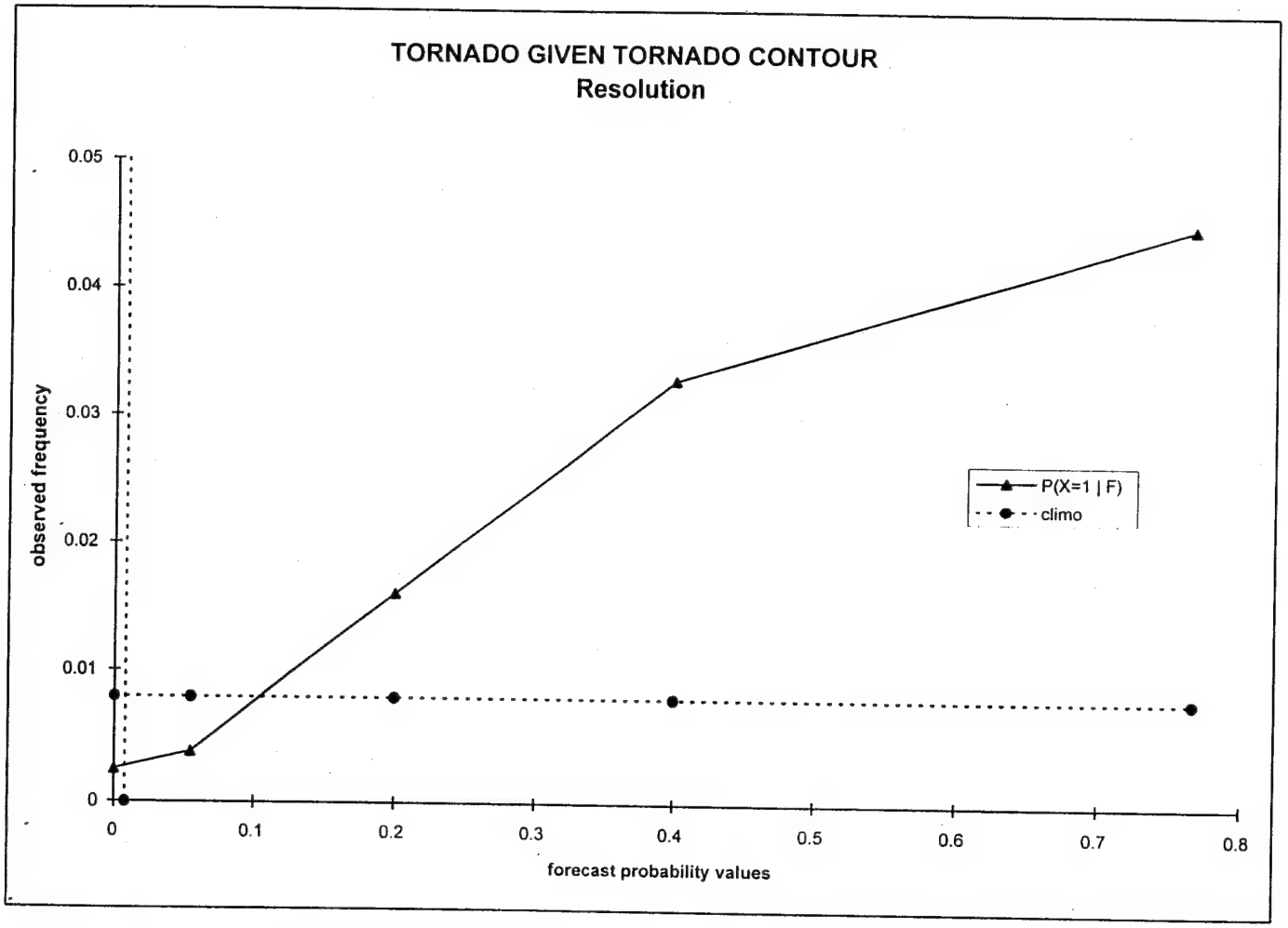


Figure 10b. Resolution diagram for tornado given tornado contour forecasts.

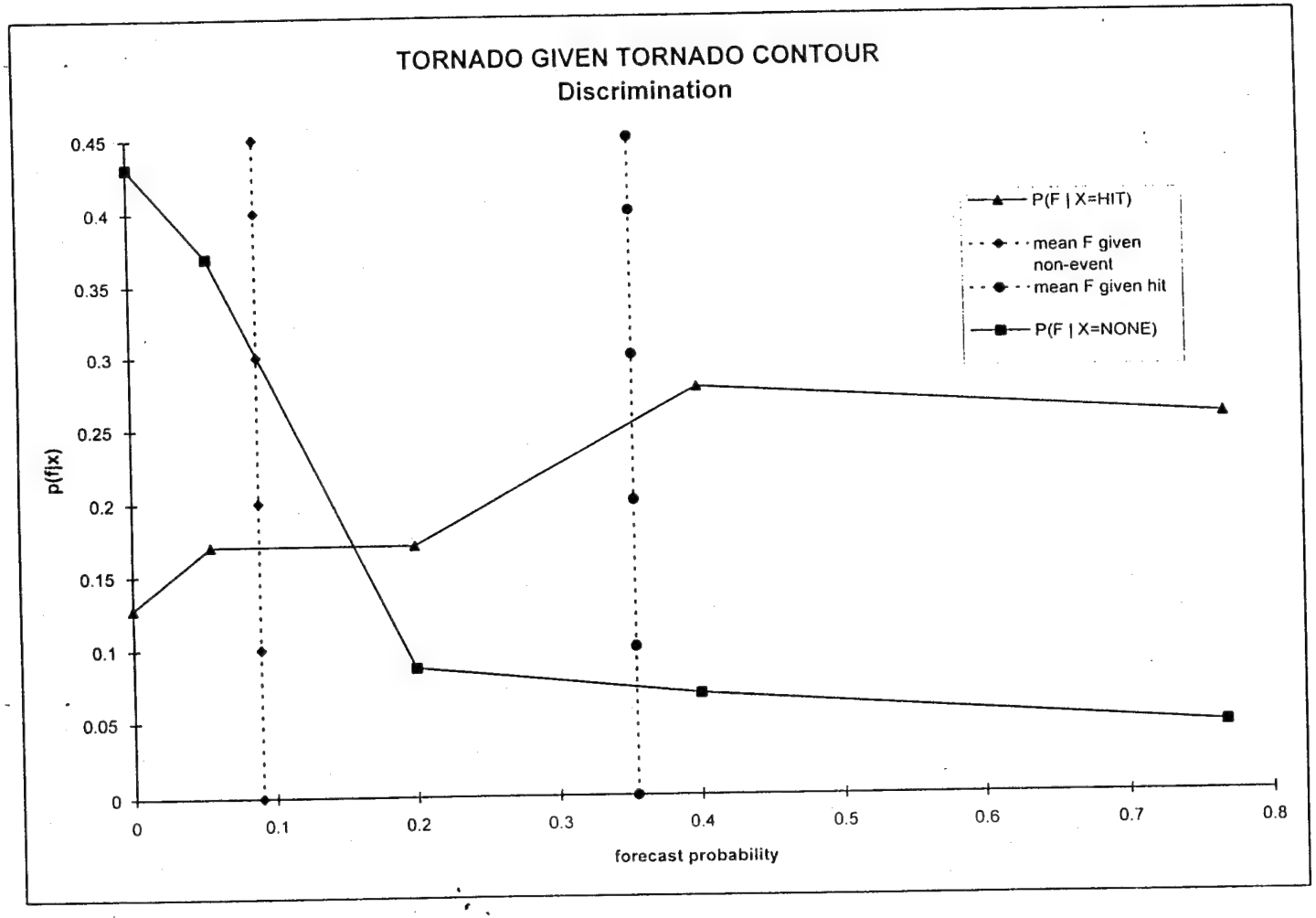


Figure 10c. Discrimination diagram for tornado given tornado contour forecasts.

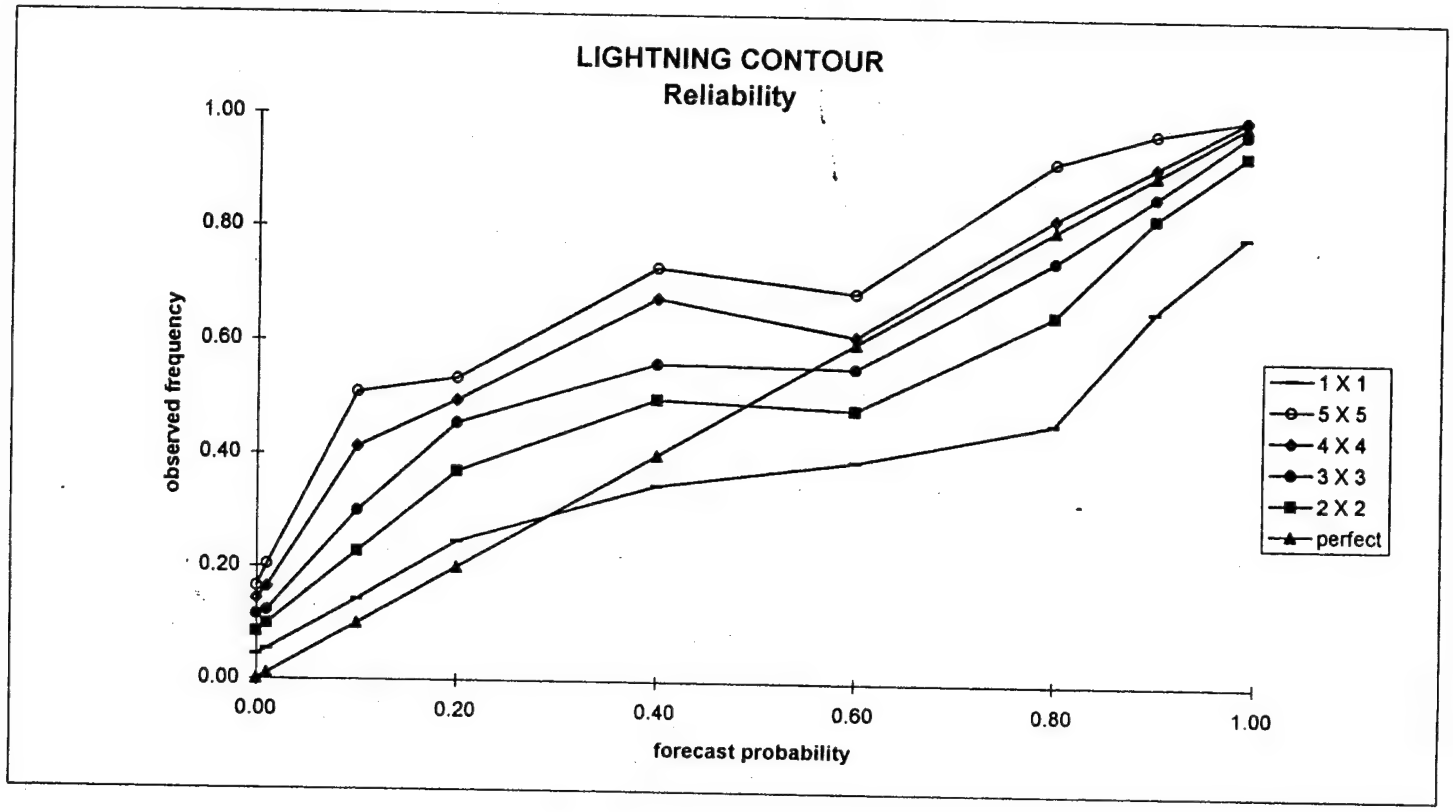


Figure 11a. Reliability diagram for different grid sizes for lightning contour forecasts.

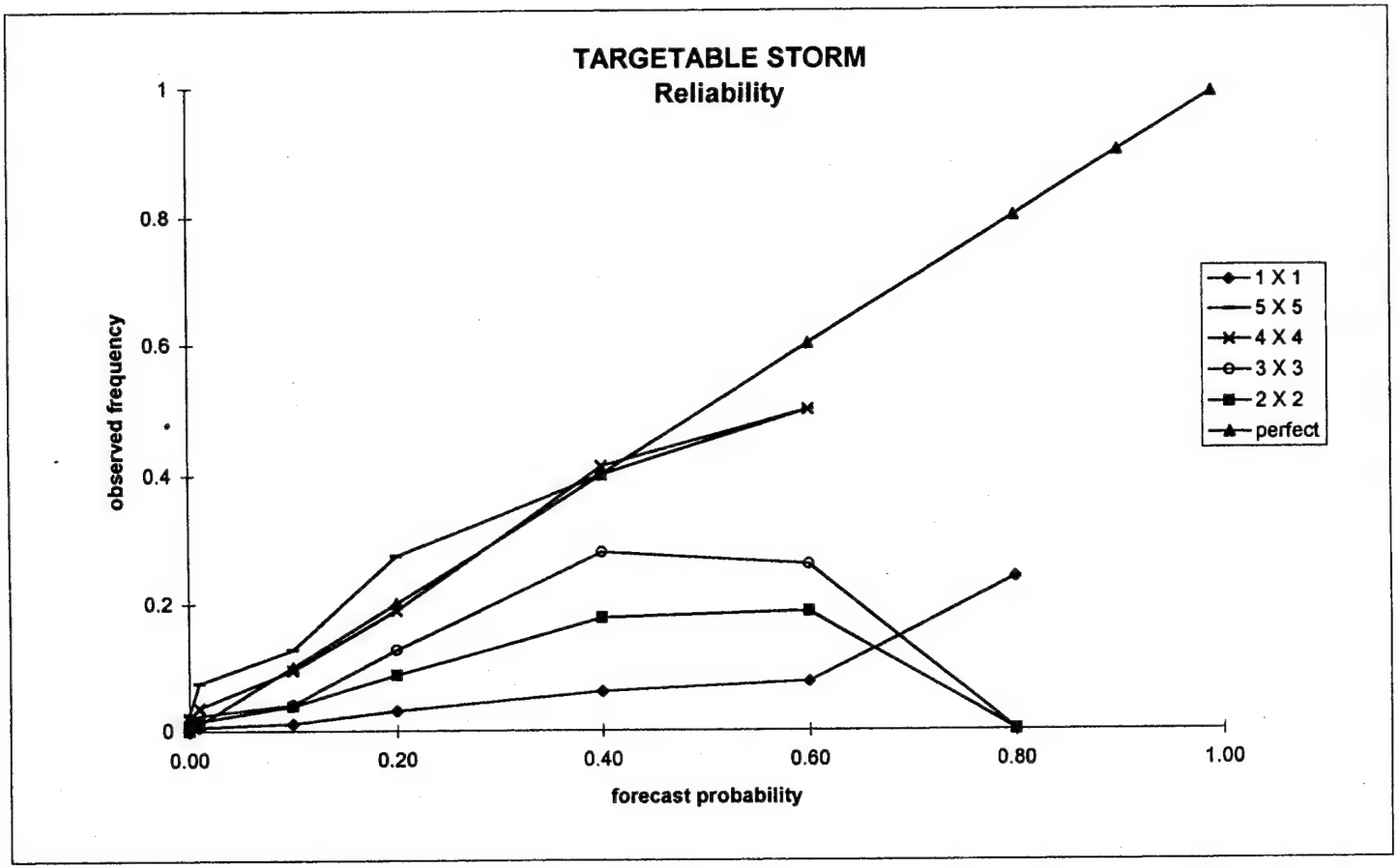


Figure 11b. As in Fig. 11a, except for targetable storm contour forecasts.

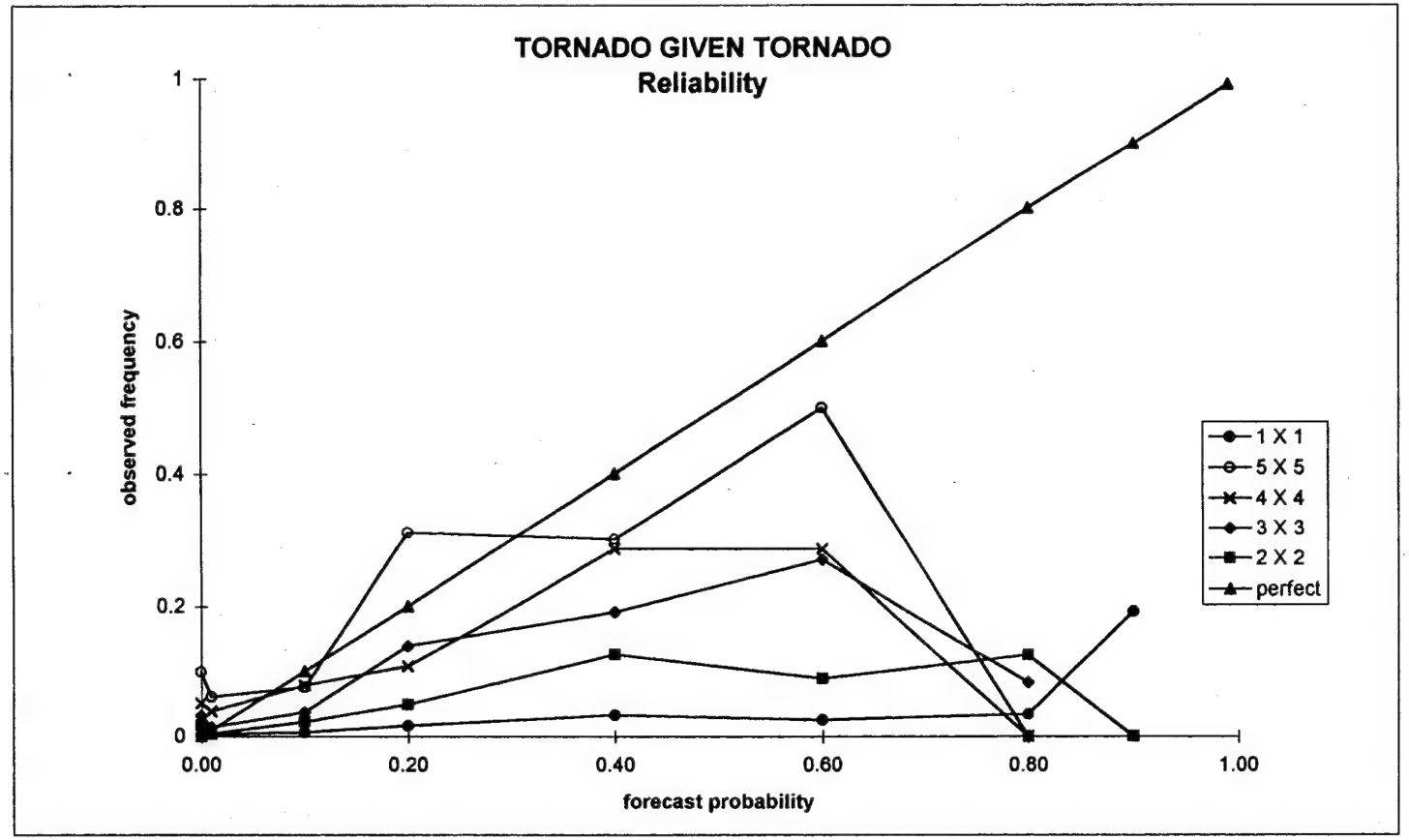


Figure 11c. As in Fig. 11a, except for tornado given tornado contour forecasts.

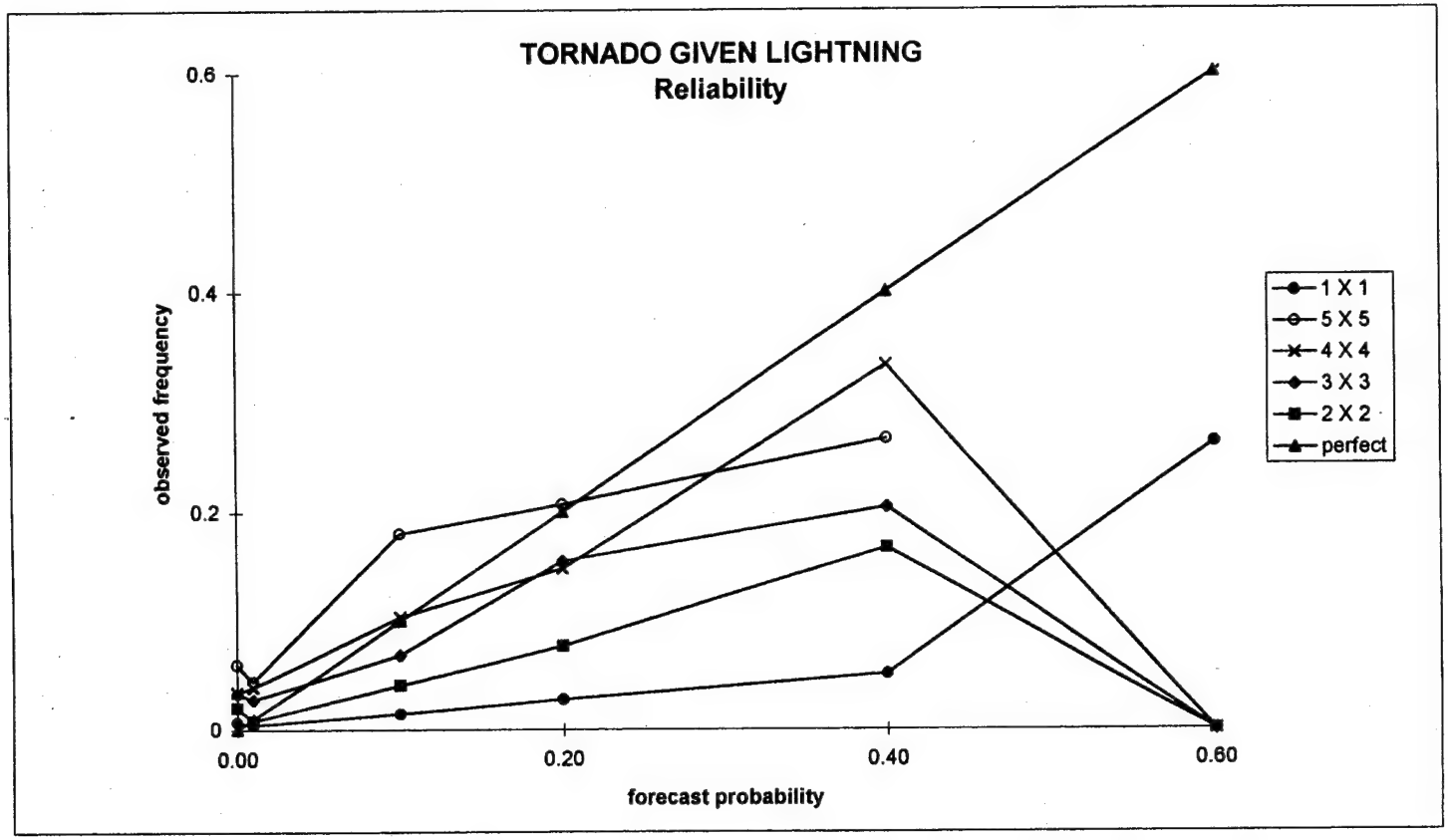


Figure 11d. As in Fig. 11a, except for tornado given lightning contour forecasts.

CHAPTER 5
CONCLUSIONS AND DISCUSSION

The VORTEX '94 experiment provided a unique opportunity for participating forecasters to gain experience at using probabilities to forecast lightning, severe weather, targetable storms, and tornadoes. Though many of the forecasters had made severe weather forecasts previously, none had forecast specifically for the VORTEX area, and certainly not with the time and space specificity required of the contoured probability forecasts. The forecasts were designed both to support field operations, and to test new experimental forecasting techniques.

Without verification, there is no feedback to the forecasters on the individual strengths and weaknesses of their forecasts and no improvement in the forecasts in time. A verification shows where the forecasts can be improved to better serve the users of the forecast. There are many possible ways to evaluate forecast quality and this paper has considered both the traditional **measures-oriented approach**, which evaluates a limited set of measures of accuracy and skill, and the **distributions-oriented approach**, which considers the joint distribution of forecasts and observations and the relationships that can be found between the two. A measures-oriented approach falls short because valuable information is lost when the information in the multi-dimensional contingency tables is reduced to a few summary measures. Using the conditional and marginal distributions associated with the distributions-oriented approach allows us to identify the particularly strong and weak areas of our forecasts.

Traditional measures including the Brier score, skill score, and bias of the forecasts were examined for this set of forecasts. The Brier score told us which

forecasts were the most accurate (lightning), but couldn't identify which parts of the forecast range were more or less accurate than others. At first glance, the Brier scores for the contour forecasts appeared to be very good, but the scores were dominated by the large number of correct null forecasts. The skill score told us which forecasts were the most skillful (severe weather) but couldn't identify which parts of the forecast range were more or less skillful than others. The bias of the forecasts indicated which phenomena were over- or underforecast but couldn't identify which parts of the forecast range were more or less biased. The area forecasts were all underforecast, which means that the forecasters are underestimating the frequency with which these phenomena occur in this relatively large area. The contour forecasts were all overforecast, with the forecasters using probabilities that were too high over areas that were too large considering the rarity of the events and the verification method (by MDR boxes). The verification with the measures-oriented approach does evaluate the overall accuracy and skill of the forecasts and identifies any biases, but it doesn't indicate specifically **where** the strengths and weaknesses lie. It is because of these weaknesses in the measures-oriented approach that we turned to the distributions-oriented approach to verify these forecasts.

The conditional and marginal distributions were examined with the $p(f \mid x)$ related to the discrimination, and the $p(x \mid f)$ related to the reliability and resolution. Summary measures were computed for each of these three quality indicators to identify the best and worst of each type, but these measures also didn't work well for contour forecasts. Explaining why one forecast excels over another is best done graphically. Graphical presentation of these three quality indicators furthered our understanding of the physical relationships between forecasts and observations by examining **where** the forecasts were more reliable, had better resolution, or were more

discriminatory. For a summary of the best and worst forecasts for each of these quality indicators, see Table 8.

The reliability diagram was helpful in identifying areas in the forecast range where the observed frequency didn't match the forecast probability. Reliability can be improved by producing unbiased forecasts. For VORTEX '94, this means using higher probabilities more often for the area forecasts and lower probabilities over larger areas for the contour forecasts. The most skillful forecasts were also the forecasts with the best resolution. The resolution diagram identified where in the forecast range the forecasts were more or less skillful. Though tornado forecasts had the best reliability, they had the worst resolution because they stuck closer to climatology than any other forecast. This is simply an acknowledgment of the inability of the forecasters to forecast tornadoes with as much confidence as other phenomena. The contour forecasts showed very little resolution, but were at least able to improve slightly over climatology. To improve the resolution of the forecasts, then, it would be helpful to know the climatology of the events being forecast and then, try to beat climatology. During VORTEX '94, the lack of knowledge of the climatology of these events hampered the forecasters ability to pick appropriate probabilities. Having this information would help the reliability, resolution, and discrimination considerably. The discrimination diagram indicated where the forecasts were able to discriminate among the different observations. Ideally, different forecast distributions follow events than those that follow non-events. To improve the discrimination of the forecasts, the forecasts should be close to one when events occur and close to zero when non-events occur. This will, of course, improve as the state-of-the-art in forecasting these phenomena improves. The discrimination got worse for rarer events, with lightning forecasts showing the best discrimination and tornado forecasts the

worst. For contour forecasts, all forecasts above 40% showed virtually no discrimination, but at least the forecasters used higher probabilities more often than lower probabilities on event days when below 40%. It's questionable whether probabilities above 40% should be used in forecasting these rare events. VORTEX '94 forecasters probably shouldn't have used probabilities this high for the contour forecasts.

There were flaws in the way that this experiment was designed that may have set the forecasters up for failure. First, the grid size used in this experiment is apparently too small. It was shown that the reliability depended on the grid size, so experiments need to be done with larger grid sizes. Second, the verification system should have been in place before the experiment started to provide timely feedback to the forecasters so that they could make adjustments. Finally, it's important that the forecasters know the climatology of the events they are forecasting so they can choose appropriate probabilities to use for the area they are forecasting. Future verifications should use the lessons learned here to improve the way verification systems are designed.

There are several other things that should be done for this set of forecasts. It was noted for the area forecasts that there appeared to be a relationship between the percentage of the area hit and the forecast probability. Although this study didn't consider areal coverage in verifying the Day-1 and Day-2 forecasts, future verifications may want to treat the forecast as an expected areal coverage forecast and verify it as such. The contour forecasts were designed to locate where the events were going to occur so, the phase errors between the forecast and observed events need to be verified. There were other parameters forecast for this experiment, including storm

motion and initiation time forecasts that have yet to be verified and, the verification of VORTEX '95 forecasts has yet to be accomplished also.

This verification has implications far beyond just verifying for a field experiment. It has broad applications to operational forecasters in the National Weather Service, especially at the Storm Prediction Center, and the military to just name a few. Using a distributions-oriented approach can provide specific feedback to the forecasters on areas where they can make technical improvements to their forecasts. For this feedback to be effective, it must be timely, and the forecasters need to understand how the vast information offered by this approach can help them. With appropriate feedback, the forecasters can distribute their forecasts in such a way to show real improvements in the reliability, resolution, and discrimination of their forecasts. These are only technical improvements, however, and we've said nothing about the meteorology behind the good and bad forecasts. This verification won't be complete until we "close the loop" by studying the meteorological reasons behind the particularly good and bad forecasts. Further studies of this kind can only serve to improve the quality of our forecasts and our verification techniques. If there is any concern with the quality of our forecasts, then verification must be done.

Table 8. Summary of best and worst forecasts.

	Day-1			Day-2			Contour	
Measure	best	worst		best	worst		best	worst
reliability	tornado	lightning		tornado	lightning		lightning	tgt
resolution	severe	tornado		targ. storm	tornado		lightning	tgt
discrimination 1	lightning	tornado		lightning	tornado		lightning	tgl
discrimination 2	lightning	tornado		lightning	tornado		lightning	tgl

BIBLIOGRAPHY

Anthony, R. W., 1990: Trends in severe local storm watch forecast performance at the NSSFC, 1978-1989. *Preprints, 16th Conf. on Severe Local Storms*, Amer. Meteor. Soc., pp 281-287

Bosart, L. F. and M. G. Landin, 1994: An assessment of thunderstorm probability forecasting skill. *Wea. Forecasting*, **9**, pp 522-531

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, pp 1-3

Brooks, H. E., C. A. III Doswell, and L. J. Wicker, 1993: STORMTIPE: A forecasting experiment using a three-dimensional cloud model. *Wea. Forecasting,* **8**, pp 352-362

_____and C. A. III Doswell, 1996: A comparison of measures-oriented and distributions-oriented approaches to forecast verification. *Wea. Forecasting,* in press

Carter, G. M. and P. D. Polger, 1986: A 20-year summary of National Weather Service verification results for temperature and precipitation. NOAA Tech. Memo. NWS FCST-31, NTIS Accession No. PB88 235353/AS, 50 pp.

Crowther, H. G. and J. T. Halmstad, 1994: Severe local storm warning verification for 1994. NOAA Tech. Memo. NSSFC-41, National Severe Storms Forecast Center, Kansas City, MO, June, 1995

Dagostaro, V. J., C. L. Alex, G. M. Carter, J. P. Dallavalle, and P. D. Polger, 1989: Evolution of the NWS national verification system: Past, present, and future. *Preprints, 11th Conf. on Probability and Statistics in Atmospheric Sciences,* Monterey, Amer. Meteor. Soc., pp J41-J46

_____and J. P. Dallavalle, 1991: AFOS-era verification of guidance and local aviation/public weather forecasts--No. 11 (October 1988 - March 1989). TDL Office Note 91-2, NWS, NOAA, U. S. Department of Commerce, 64 pp.

_____and _____, 1995: AFOS-era verification of guidance and local aviation/public weather forecasts--No. 20 (April 1993 - Septemper 1993). TDL Office Note 95-1, National Weather Service, NOAA, U. S. Department of Commerce, 50 pp

_____, _____, M. D. Miller, and V. C. Southall, 1995: AFOS-era verification of guidance and local aviation/public weather forecasts--No. 21 (October 1993 - March 1994). TDL Office Note 95-2, National Weather Service, NOAA, U. S. Department of Commerce, 52 pp.

Doswell, C. A. III, R. A. Maddox, and C. F. Chappell, 1986: Fundamental considerations in forecasting for field experiments. *Preprints, 11th Conf. on Weather Forecasting and Analysis.* Amer. Meteor. Soc., pp 353-358

_____ and J. A. Flueck, 1989: Forecasting and verifying in a field research project: DOPLIGHT '87. *Wea. Forecasting,* **4**, pp 97-109

_____, R. Davies-Jones, and D. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Wea. Forecasting,* **5**, pp 576-585

Galway, J. G., 1967: SELS forecast verification, 1952-1966. *Preprints, 5th Conf. on Severe Local Storms*, St. Louis, MO, Amer. Meteor. Soc., pp 140-145

Jincai, D., C. A. III Doswell, D. W. Burgess, M. P. Foster, and M. L. Branick, 1992: Verification of mesoscale forecasts made during MAP '88 and MAP '89. *Wea. Forecasting,* **7**, pp 468-479

Murphy, A. H., 1979: On the evaluation of point precipitation probability forecasts in terms of areal coverage. *Mon. Wea. Rev.,* **106**, pp 1680-1686

_____and R. L. Winkler, 1982: Subjective probabilistic tornado forecasts: Some experimental results. *Mon. Wea. Rev.*, **110**, pp 1288-1297

_____and H. Daan, 1984: Impacts of feedback and experience on the quality of subjective probability forecasts: Comparison of results from the first and second years of the Zierikzee experiment. *Mon. Wea. Rev.*, **112**, pp 413-423

_____, W. -R. Hsu, R. L. Winkler, and D. S. Wilks, 1985: The use of probabilities in subjective Quantitative Precipitation Forecasts: Some experimental results, *Mon. Wea. Rev.,* **113**, pp 2075-2089

_____and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.,* **115**, pp 1330-1338*

_____, B. G. Brown, and Y. -S. Chen, 1989: Diagnostic verification of temperature forecasts. *Wea. Forecasting,* **4**, pp 485-501

_____, 1991: Forecast verification: It's complexity and dimensionality. *Mon. Wea. Review,* **119**, pp 1590-1601

_____, 1991: Probabilities, odds, and forecasts of rare events. *Wea. Forecasting,* **6**, pp 302-307

_____, 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting,* **8**, pp 281-293

_____, 1996: The Finley affair: A signal event in the history of forecast verification. *Wea. Forecasting,* **11**, pp 3-20

National Weather Service, 1982: *National Verification Plan.* NOAA, U. S. Department of Commerce, 81 pp.

Pearson, A. D. and S. J. Weiss, 1979: Some trends in forecast skill at the National Severe Storms Forecast Center: 1967-1977, *Bull. Amer. Meteor. Soc.*, **60**, pp 319-326

Rasmussen, E. N., J. M. Straka, R. Davies-Jones, C. A. III Doswell, F. H. Carr, M. D. Eilts, and D. R. MacGorman, 1994: Verification of the origins of rotation in tornadoes experiment: VORTEX. *Bull. Amer. Meteor. Soc.,* **75**, pp 995-1006

Sanders, F., 1963: On subjective probability forecasting. *J. Appl. Meteor.*, **2**, pp 191-201

Weiss, S. J., D. L. Kelly, and J. T. Schaefer, 1980: New objective verification techniques at the National Severe Storms Forecast Center. *Preprints, 8th Conf. on Weather Forecasting and Analysis,* Denver, CO, Amer. Meteor. Soc., pp 140-145

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences,* R. Dmowska and J. R. Holton, Eds., Academic Press, pp 233-283

APPENDIX A

VORTEX FORECAST INFORMATION

VORTEX 94 FORECASTERS

Lead Forecasters

Mike Branick (WSFO/EFF OUN)
Don Burgess (OSF)
Chuck Doswell (NSSL)
Jack Hales (NSSFC/SELS)
Bob Johns (NSSFC/SELS)
Larry Ruthi (WSFO OUN)
Steve Weiss (NSSFC/SELS)

Assistant Forecasters

Phil Bothwell (SPC, Norman)
Harold Brooks (NSSL)
Dennis Dudley (AES, Winnipeg, Man., Canada)
Mike Leduc (AES, King City, Ont., Canada)
Roger Edwards (NSSFC/SELS)
Paul Janish (NSSL/EFF OUN)
John Cortinas (NSSL/CIMMS)

Figure A1. List of VORTEX '94 forecasters.

DATE: 04 / 01 /1994

ISSUE TIME: 0900 LT

FORECASTER: DOSWELL

PROBABILITY OF OCCURRENCE WITHIN VORTEX FORECAST AREA
(Values: 0,2,5,10,20,30,40,50,60,70,80,90,95,98,100%):

	DAY 1	*DAY 2*
CONVECTION (LIGHTNING)	[80]	[90]
SEVERE CONVECTION	[60]	[80]
TORNADOES	[20]	[60]
TARGETABLE STORM	[30]	[70]

FORECAST INITIATION TIMES (CDT hour - Targetable storms only):

FIRST LIGHTNING	[1530]	[1400]
FIRST SEVERE REPORT	[1600]	[1430]
FIRST MESO/TORNADO	[1630]	[1530]

FORECAST STORM MOTION (dir/speed[*kt*] - Targetable storms only):

PRE-SUPERCELL	[250/25]	[260/30]
SUPERCELL	[300/20]	[280/25]

GENERAL LOCATION OF TARGET AREA - DAY 1: [Central TX, E of LBB-MAF]

GENERAL LOCATION OF TARGET AREA - DAY 2: [North Tx, Srn OK]

DISCUSSION/COMMENTS:

Significant trough dropping SEwd, ETA progs slowing it considerably! Mstr incrg into central TX thru DRT Timing is critical question If trof slows down, mstr will have more time to return 700-500 Δt's already 20+ chd of trof. Expect good chc of svr tda in TX. With sys slowing down tomorrow has real potential. probably near Red River east of CDS as trof increases helicity (expect mstr - instblty to be in place then.) Another problem for tomorrow is effect of convection overnite.

Figure A2. Sample area forecast.

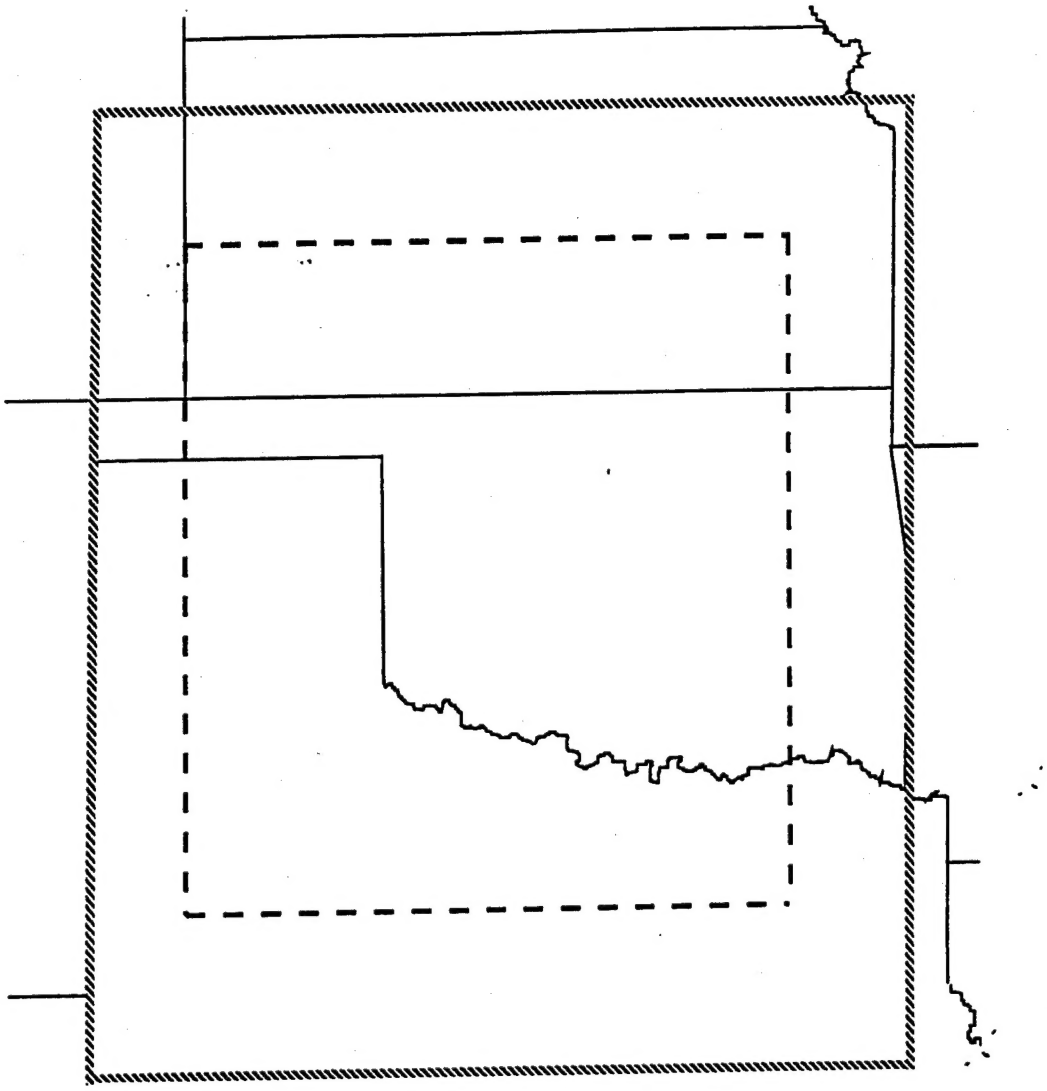


Figure A3. VORTEX operational area (inner rectangle) and forecast area (outer rectangle).

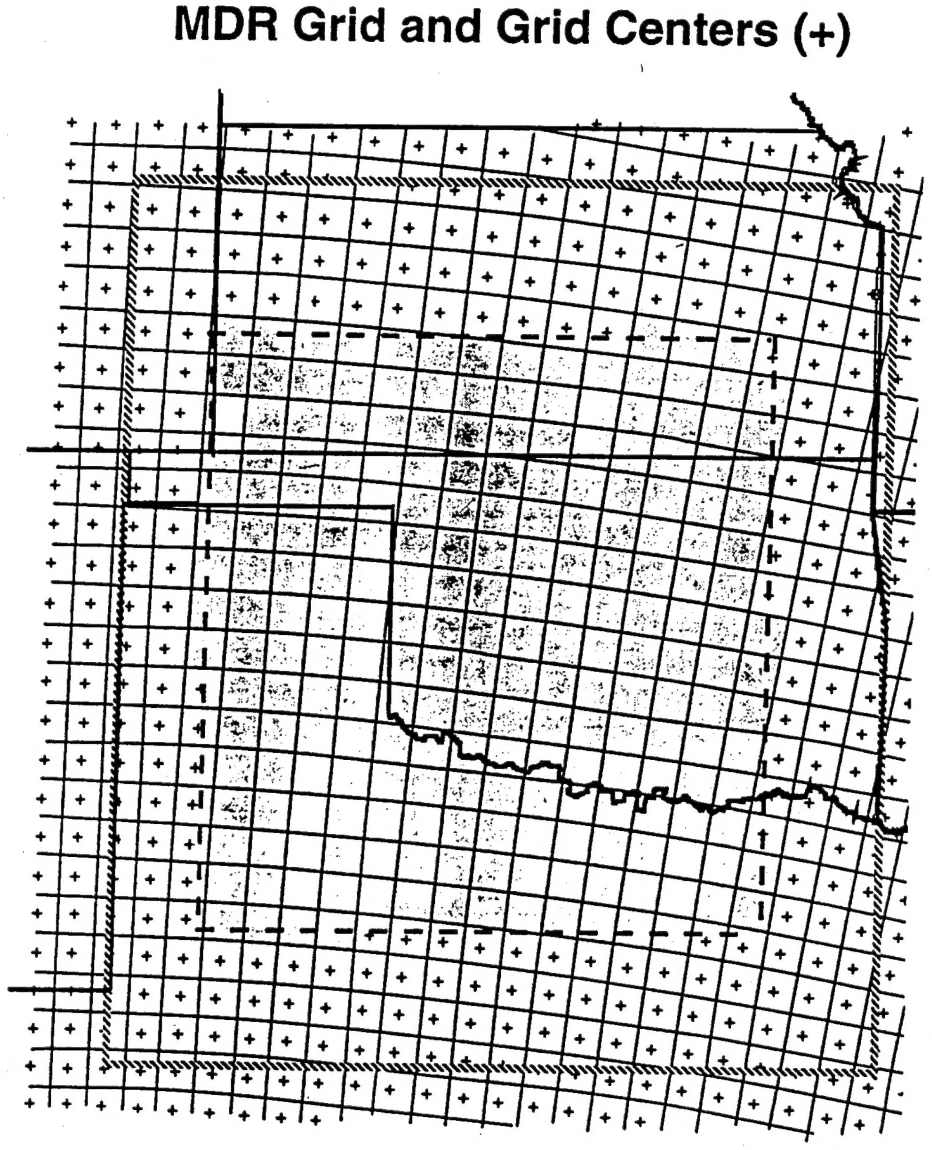


Figure A4. MDR (manually-digitized) radar grid over VORTEX area.

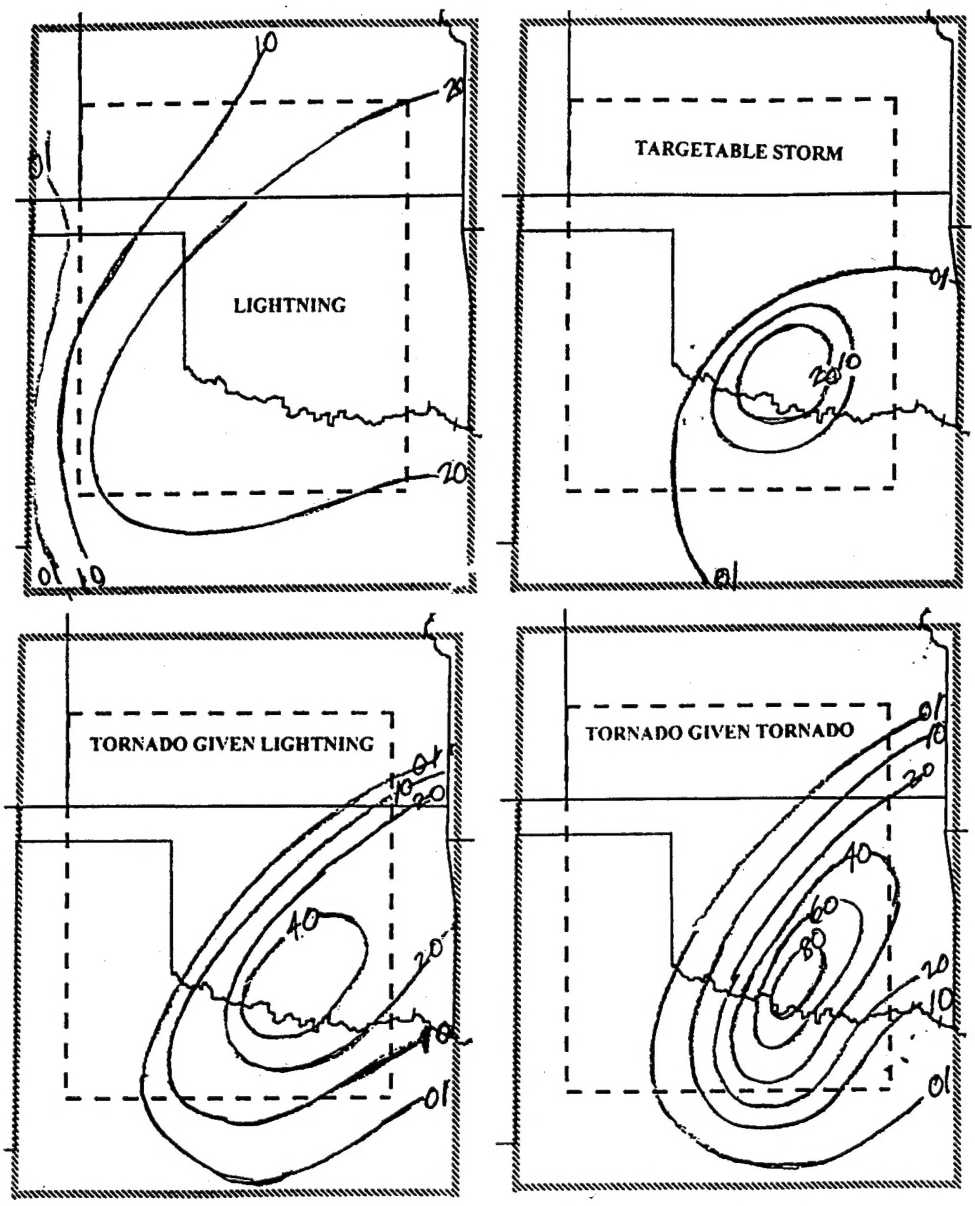


Figure A5. Sample contour forecast.